Order-Based Dependent Dirichlet Processes

J.E. Griffin and M.F.J. Steel*

Abstract

In this paper we propose a new framework for Bayesian nonparametric modelling with continuous covariates. In particular, we allow the nonparametric distribution to depend on covariates through ordering the random variables building the weights in the stick-breaking representation. We focus mostly on the class of random distributions which induces a Dirichlet process at each covariate value. We derive the correlation between distributions at different covariate values, and use a point process to implement a practically useful type of ordering. Two main constructions with analytically known correlation structures are proposed. Practical and efficient computational methods are introduced. We apply our framework, though mixtures of these processes, to regression modelling, the modelling of stochastic volatility in time series data and spatial geostatistical modelling.

Keywords: Bayesian nonparametrics, Markov chain Monte Carlo, Nonparametric Regression, Spatial Modelling, Stick-breaking Prior, Volatility Modelling.

1 Introduction

Bayesian nonparametric method have become increasingly popular in empirical studies. The Dirichlet process (Ferguson 1973) has been the dominant mechanism used as the prior for the unknown distribution in the model specification. Some recent examples include applications in econometrics (Chib and Hamilton 2002; Hirano 2002), medicine (Kottas *et al.* 2002), health (O'Hagan and Stevens 2003), auditing (Laws and O'Hagan 2002), animal breeding (van der Merwe and Pretorius 2003), survival analysis

*Jim Griffin is Lecturer, Department of Statistics, University of Warwick, CV4 7AL, U.K. (Email: J.E.Griffin@warwick.ac.uk) and Mark Steel is Professor, Department of Statistics, University of Warwick, Coventry, CV4 7AL, U.K. (Email: M.F.Steel@stats.warwick.ac.uk). Both authors were affiliated with the Institute of Mathematics, Statistics and Actuarial Science, University of Kent at Canterbury during the early part of this research. Jim Griffin acknowledges research support from The Nuffield Foundation grant NUF-NAL/00728. We would like to thank Andy Hone for his contribution to some calculations and we are grateful to Carmen Fernández and Steve MacEachern for helpful discussions and to two referees and the Associate Editor for insightful comments.

(Doss and Huffer 2003), directional data (Ghosh *et al.* 2003), meta analysis (Chung *et al.* 2002), genetics (Medvedovic and Sivaganesan 2002) and density estimation (Hansen and Lauritzen 2002). However, modelling the relationship between covariates and the unknown distribution cannot be achieved directly using the Dirichlet process described by Ferguson.

Therefore, an active area of research is extending these methods to a wider class of models where the unknown distribution depends on covariates. If the covariates have a finite number of levels the Product of Dirichlet processes model introduced by Cifarelli and Regazzini (1978) allows the modelling of dependent distributions. Dependence is introduced through the use of a parametric regression model as the centring distribution of independent Dirichlet processes at each level of the covariates. These method have recently been applied to problems in biostatistics (Carota and Parmigiani 2002), econometrics (Griffin and Steel 2004) and survival analysis (Guidici *et al.* 2003) and a similar idea was proposed in Mallick and Walker (1997). In the present paper we focus on introducing dependence on continuous covariates. Other approaches to this problem exist in the literature. Müller and Rosner (1998) propose including the covariates in the nonparametric distribution and focusing on the conditional given the covariates only. Since this implies leaving out a factor in the likelihood, Müller *et al.* (2004) change the prior on the process to counteract this fact. Finally, the method described by MacEachern *et al.* (2001) is closest to the approach developed here, as both approaches start from the Sethuraman (1994) representation, mentioned in the following subsection.

Here we introduce dependence in nonparametric distributions by making the weights in the Sethuraman representation dependent on the covariates. Each weight is a transformation of i.i.d. random variables. The way we implement the dependence is by inducing an ordering π of these random variables at each covariate value such that distributions for similar covariates values will be associated with similar orderings and, thus, be close. At any covariate value, the random distribution will be a so-called stick-breaking prior. We focus on the special case where we choose the Dirichlet process for this stick-breaking prior, and we shall call the induced class of processes Order-Based Dependent Dirichlet Processes, shortened to πDDP's.

We derive theoretical properties, such as the correlation between distributions at different covariate values, and use a point process to implement a practically useful type of ordering. Two main constructions with analytically known correlation structures are proposed. Practical computational methods are introduced, using Markov chain Monte Carlo (MCMC) methods. We control the truncation error in an intuitive fashion through truncation of the point process and we use sequential allocation as an efficient way to avoid the sampler getting stuck in local modes. We apply our basic framework, though mixtures of πDDP's, in three quite different settings. We use it for curve fitting, the modelling of stochastic volatility in time series data and spatial geostatistical modelling.

Subsection 1.1 describes stick-breaking priors, while Section 2 introduces the ideas underlying πDDP's and their practical implementation. Section 3 briefly discusses mixtures of these processes, and Section 4 concerns elicitation of the prior. Computational issues are dealt with in Section 5 and Section 6 describes the three applications. The final section concludes.

Proofs will be grouped in Appendix A without explicit mention in the text.

1.1 Stick-breaking priors

The idea of defining random distributions through stick-breaking construction is developed in Pitman (1996) where its uses in several areas of application are reviewed. The class is discussed by Ishwaran and James (2001) as a prior distribution in nonparametric problems. A random distribution, F, has a stick-breaking prior if

$$F \stackrel{d}{=} \sum_{i=1}^{N} p_i \delta_{\boldsymbol{\theta}_i}, \tag{1}$$

where δ_z denotes a Dirac measure at z, $p_i = V_i \prod_{j<i}(1 - V_j)$ where $V_1, \ldots, V_N$ are independent with $V_k \sim \text{Beta}(a_k, b_k)$ and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ are independent draws from a distribution H. Conventionally, only models with an infinite representation are referred to as nonparametric (see *e.g.* Bernardo and Smith, 1994, p.228). Ishwaran and James (2001) give the following condition to determine if the distribution is well-defined for $N = \infty$

$$\sum_{k=1}^{\infty} p_k = 1 \text{ a.s.} \iff \sum_{k=1}^{\infty} \text{E}(\log(1 - V_k)) = -\infty.$$

For finite N the condition $\sum_{k=1}^{N} p_k = 1$ is satisfied if $V_N = 1$ so that $p_N = \prod_{j<N}(1 - V_j)$. For $N = \infty$ several interesting processes fall into this class:

1. The Dirichlet process prior (Ferguson 1973) characterised by MH, where M is a positive scalar, arises when V_i follows a Beta$(1, M)$ for all i. This was established by Sethuraman (1994).
2. The Pitman-Yor process occurs if V_i follows a Beta$(1 - a, b + ai)$ with $0 \leq a < 1$ and $b > -a$. As special cases we can identify the Dirichlet process for $a = 0$ and the stable law when $b = 0$.

This representation will provide the basis for our development of dependent probability measures and, in particular, the development of a dependent Dirichlet process. We will refer to the $\boldsymbol{\theta}_i$'s as locations and the V_i's as masses.

2 Dependent Dirichlet Processes

2.1 General construction

A dependent Dirichlet process is a stochastic process defined on the space of probability measures over a domain, indexed by time, space or a selection of other covariates in such a way that the marginal distribution at any point in the domain follows a Dirichlet process. This problem has received little attention in the Bayesian literature. Some recent work follows MacEachern (1999). The latter paper considers the possibility of allowing the masses, V, or the locations, $\boldsymbol{\theta}$, of the atoms to follow a stochastic process defined over the domain. An important constraint imposed by the definition of the Dirichlet process is that the processes for each element of either $\boldsymbol{\theta}$ or V must be independent. The work of MacEachern and coauthors concentrates on the "single-p" model where only the locations, $\boldsymbol{\theta}$, follow stochastic processes. An application to spatial modelling is further developed in Gelfand *et al.* (2004)

by allowing the locations $\boldsymbol{\theta}$ to be drawn from a random field (a Gaussian process). The same method to induce dependence is used in De Iorio *et al.* (2004) to achieve an analysis of variance (ANOVA)-type structure.

In general, such approaches which allow only values of $\boldsymbol{\theta}$ to depend on the covariates are subject to certain problems. In particular, MacEachern notes that the distribution of F can then be expressed as a mixture of Dirichlet processes. The posterior process will have an updated mass parameter $M + n$, where n is the sample size, *at all values of the index*. This latter fact is counterintuitive, in our view. A useful property would rather be that the process returns to the prior distribution (with mass parameter M) at points in the domain "far" from the observed data. This seems a major shortcoming of these single-p models for general spaces.

In contrast to the models described above, the processes developed in this paper allow the values of the weights p_i in (1) to change over the domain of the covariates. For ease of presentation it will be assumed that each location, $\boldsymbol{\theta}_i$, does not depend on the covariates. However, the ideas that are developed could be extended to allow for the introduction of dependence through the locations (*i.e.* drawn from independent stochastic processes). MacEachern (2000) has some useful results in this direction.

Definition 1 *An Order-based Dependent Stick-Breaking Prior is defined on a space D by a sequence $\{a_k, b_k\}$, centring distribution H and a stochastic process $\{\pi(\mathbf{x})\}_{\mathbf{x}\in D}$ for which:*

1. *$\{\pi_1(\mathbf{x}), \ldots, \pi_{n(\mathbf{x})}(\mathbf{x})\} \subseteq \{1, \ldots, N\}$ for some $n(\mathbf{x}) \leq N$.*
2. *$\pi_i(\mathbf{x}) = \pi_j(\mathbf{x})$ if and only if $i = j$.*

Random variables $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ and $V_1, \ldots, V_{N-1}$ are all independent, $\boldsymbol{\theta}_k \sim H$ and $V_k \sim Beta(a_k, b_k)$. The distribution at a point $\mathbf{x} \in D$ is defined by

$$F_{\mathbf{x}} \stackrel{d}{=} \sum_{i=1}^{n(\mathbf{x})} p_i(\mathbf{x}) \delta_{\boldsymbol{\theta}_{\pi_i(\mathbf{x})}}$$

$$p_i(\mathbf{x}) = V_{\pi_i(\mathbf{x})} \prod_{j<i} (1 - V_{\pi_j(\mathbf{x})}),$$

and for finite $n(\mathbf{x})$

$$p_{n(\mathbf{x})}(\mathbf{x}) = \prod_{j<n(\mathbf{x})} (1 - V_{\pi_j(\mathbf{x})}).$$

We will refer to $\boldsymbol{\pi}(\mathbf{x}) = (\pi_1(\mathbf{x}), \ldots, \pi_{n(\mathbf{x})}(\mathbf{x}))$ as the ordering at $\mathbf{x}$.

The stick-breaking prior in Subsection 1.1 is recovered for any given $\mathbf{x} \in D$. We obtain the same distribution over the entire space D if $\pi_i(\mathbf{x}) = i$ for all $\mathbf{x} \in D$ and $i = 1, \ldots, N$. As a more interesting example, the stochastic process $\pi(\mathbf{x})$ could be defined on the space of permutations of $\{1, \ldots, N\}$ (*i.e.* $n(\mathbf{x}) = N$ for all $\mathbf{x} \in D$), allowing $F_{\mathbf{x}}$ to change with $\mathbf{x}$. However, the definition allows the stochastic process to be defined on more general structures. In particular, some elements of the ordering at a given point need not appear in the ordering at another point. An example of such a process is given in Subsection 2.2.2. In general, this defines a wide class of dependent distributions, both parametric (finite

N) and nonparametric (infinite N). Usually, we will be interested in $N = \infty$ so that $F_{\mathbf{x}}$ can follow a Dirichlet process. However, it is not easy to define stochastic processes $\pi(\mathbf{x})$ for infinite N. Therefore, we focus our attention on specific constructions for the stochastic process in this case.

The prior distribution for $F_{\mathbf{x}}$ inherits some properties of stick-breaking priors. For example, the first moment measure is

$$\mathrm{E}[F_{\mathbf{x}}(B)] = \mathrm{E}\left[\sum_{i=1}^{n(\mathbf{x})} p_i(\mathbf{x})\delta_{\boldsymbol{\theta}_{\pi_i(\mathbf{x})}}(B)\right] = \mathrm{E}\left[\sum_{i=1}^{n(\mathbf{x})} p_i(\mathbf{x})\right]\mathrm{E}\left[\delta_{\boldsymbol{\theta}_{\pi_i(\mathbf{x})}}(B)\right] = H(B),$$

and

$$\begin{aligned}\mathrm{Var}[F_{\mathbf{x}}(B)|\boldsymbol{\pi}(\mathbf{x})] = {} & H(B)(1-H(B)) \\ \times \sum_{i=1}^{n(\mathbf{x})} & \frac{a_{\pi_i(\mathbf{x})}\left(a_{\pi_i(\mathbf{x})}+1\right)}{\left(a_{\pi_i(\mathbf{x})}+b_{\pi_i(\mathbf{x})}\right)\left(a_{\pi_i(\mathbf{x})}+b_{\pi_i(\mathbf{x})}+1\right)}\prod_{j<i}\frac{b_{\pi_i(\mathbf{x})}\left(b_{\pi_i(\mathbf{x})}+1\right)}{\left(a_{\pi_i(\mathbf{x})}+b_{\pi_i(\mathbf{x})}\right)\left(a_{\pi_i(\mathbf{x})}+b_{\pi_i(\mathbf{x})}+1\right)}.\end{aligned}$$

In the sequel we will assume that $a_k = 1, b_k = M$ and that $N = \infty$ so that we recover a Dirichlet process at any $\mathbf{x} \in D$ if $n(\mathbf{x}) = \infty$. For the marginal variance, we then obtain

$$\mathrm{Var}\left[F_{\mathbf{x}}(B)\right] = \mathrm{E}_{\boldsymbol{\pi}(\mathbf{x})}\left[\mathrm{Var}\left[F_{\mathbf{x}}(B)|\boldsymbol{\pi}(\mathbf{x})\right]\right] = \frac{H(B)(1-H(B))}{M+1}. \tag{2}$$

The associated subclass of processes will be denoted by **Order-Based Dependent Dirichlet Processes**, abbreviated as πDDP and characterised by a mass parameter M, centring distribution H and a stochastic process $\{\pi(\mathbf{x})\}_{\mathbf{x}\in D}$.

The construction in Definition 1 is motivated by the fact that for our stick-breaking prior $\mathrm{E}[p_i(\mathbf{x})] < \mathrm{E}[p_{i-1}(\mathbf{x})]$ for any $\mathbf{x}$ and thus the influence of an atom diminishes as it gets further down the ranking (*i.e.* its order in $\pi(\mathbf{x})$ increases). This allows us to easily impose the characteristic of "localness" which can be described as follows. An important improvement over the single-p DDP models is the flexibility to allow the posterior at an index $\mathbf{x}^\star$ to tend to the prior as the distance between $\mathbf{x}^\star$ and observed indices tends to infinity if $n(\mathbf{x}) = \infty$ for all $\mathbf{x}$. If we observe $y_1, \ldots, y_n$ at indices $\mathbf{x}_1, \ldots, \mathbf{x}_n$, posterior updating can be seen as linking the observations to atoms of the distribution at each index by a new variable s_i for which $\theta_{\pi_{s_i}(\mathbf{x}_i)} = y_i$ and $P(s_i = j) = p_j(\mathbf{x}_i)$. Thus, s_i is the ranking of location y_i at index $\mathbf{x}_i$. Conditioning on $s_1, \ldots, s_n$ and $\boldsymbol{\pi}$, there will be a subset of $\{1, \ldots, n\}$, which we call $\mathcal{J}$, for which $\pi_{s_j^\star}(\mathbf{x}^\star) = \pi_{s_j}(\mathbf{x}_j)$, where the variables $s_j^\star, j = 1, \ldots, n$ denote the position of location y_j in the ordering at index $\mathbf{x}^\star$. This set $\mathcal{J}$ groups the observed locations which are in the ordering both at $\mathbf{x}_j$ and at $\mathbf{x}^\star$. Then

$$F_{\mathbf{x}^\star} \stackrel{d}{=} \sum_{i=1}^{\min\left\{\pi_{s_j^\star}(\mathbf{x}^\star)\middle|j\in\mathcal{J}\right\}} p_i(\mathbf{x}^\star)\delta_{\theta_{\pi_i(\mathbf{x}^\star)}} + \sum_{i=\min\left\{\pi_{s_j^\star}(\mathbf{x}^\star)\middle|j\in\mathcal{J}\right\}+1}^{\infty} p_i(\mathbf{x}^\star)\delta_{\theta_{\pi_i(\mathbf{x}^\star)}}.$$

The only updated elements of $\boldsymbol{\theta}$ and V in the conditional posterior will be those indexed by the elements of $\{\pi_{s_j^\star}(\mathbf{x}^\star)|j \in \mathcal{J}\}$. The first part of the sum involves random variable which have not been updated and

so we need $\min\left\{\pi_{s_j^\star}(\mathbf{x}^\star)\middle| j \in \mathcal{J}\right\}$ to increase as $\|\mathbf{x}^\star - \mathbf{x}_j\| \to \infty$ for some appropriate distance measure. Finally, if we marginalize over $s_1^\star, \ldots, s_n^\star, \boldsymbol{\pi}$, we need to check that $P\left(\min\left\{\pi_{s_j^\star}(\mathbf{x}^\star)\middle| j \in \mathcal{J}\right\} < C\right) \to 0$ as $\|\mathbf{x}^\star - \mathbf{x}_j\| \to \infty$ for all j and any finite C. This condition will hold for the constructions introduced in the sequel. Thus, our updating is made "local" by the diminishing influence of observations corresponding to indices that are increasingly far away.

The correlation between the realised distributions drawn at two points $\mathbf{x}_1$ and $\mathbf{x}_2$ is controlled by the similarity in $\pi(\mathbf{x}_1)$ and $\pi(\mathbf{x}_2)$. To make the notion of correlation between distributions more concrete we consider two related measures: the correlation between the measures of some set B at $\mathbf{x}_1$ and $\mathbf{x}_2$ (which generalizes the standard moment measures) and the correlation between points drawn at random from the two distributions. First, we consider a fixed ordering at $\mathbf{x}_1$ and $\mathbf{x}_2$, and later develop the random ordering case.

Theorem 1 *Let* $T(\mathbf{x}_1, \mathbf{x}_2) = \{k | \text{there exists } i, j \text{ such that } \pi_i(\mathbf{x}_1) = \pi_j(\mathbf{x}_2) = k\}$ *and let* $A_{lk} = \{\pi_j(\mathbf{x}_l) | j < i \text{ where } \pi_i(\mathbf{x}_l) = k\}$ *for* $k \in T(\mathbf{x}_1, \mathbf{x}_2)$. *For a given ordering* $\boldsymbol{\pi}(\mathbf{x})$, *the correlation* $Corr(F_{\mathbf{x}_1}(B), F_{\mathbf{x}_2}(B))$ *can be expressed as*

$$Corr(F_{\mathbf{x}_1}(B), F_{\mathbf{x}_2}(B)) = \frac{2}{M+2} \sum_{k \in T(\mathbf{x}_1, \mathbf{x}_2)} \left(\frac{M}{M+2}\right)^{\#S_k} \left(\frac{M}{M+1}\right)^{\#S_k'}, \tag{3}$$

where $\#A$ *is the number of distinct elements in a set* A *and*

$$S_k = A_{1k} \cap A_{2k}$$

$$S_k' = A_{1k} \cup A_{2k} - S_k.$$

If we consider the first k elements of the orderings at $\mathbf{x}_1$ and $\mathbf{x}_2$, S_k is the set of elements shared by the two orderings and S_k' are those elements that only appear in one of the orderings. For a given k, reducing $\#S_k$ by one will induce adding two elements to S_k', thus reducing the correlation, as expected.

Since the autocorrelation is not a function of B, we can think of Corr $(F_{\mathbf{x}_1}(B), F_{\mathbf{x}_2}(B))$ as "the" autocorrelation between $F_{\mathbf{x}_1}$ and $F_{\mathbf{x}_2}$, which is indicated as Corr$(F_{\mathbf{x}_1}, F_{\mathbf{x}_2})$. The correlation between two observations y_1 and y_2 drawn at random from the distributions $F_{\mathbf{x}_1}$ and $F_{\mathbf{x}_2}$ has the form

$$\begin{aligned}\text{Corr}(y_1, y_2) =& \frac{2}{(M+1)(M+2)} \sum_{k \in T(\mathbf{x}_1, \mathbf{x}_2)} \left(\frac{M}{M+2}\right)^{\#S_k} \left(\frac{M}{M+1}\right)^{\#S_k'} \\ =& \frac{1}{M+1} \text{Corr}(F_{\mathbf{x}_1}, F_{\mathbf{x}_2}).\end{aligned}$$

We can now clearly identify the separate roles of the parameters of the πDDP: the centring distribution H determines the mean of $F_{\mathbf{x}}$, the mass parameter M controls the precision and the ordering $\boldsymbol{\pi}(\mathbf{x})$ will, in combination with M, determine the dependence across the domain. In the limit as $M \to \infty$ we tend to the parametric case, and we will lose the dependence since the ordering then no longer matters (*i.e.* E$[p_i(\mathbf{x})]$ will tend to E$[p_{i-1}(\mathbf{x})]$).

Theorem 1 formalises the relationship between the orderings $\pi(\mathbf{x}_1)$ and $\pi(\mathbf{x}_2)$ and the autocorrelation between the distributions $F_{\mathbf{x}_1}$ and $F_{\mathbf{x}_2}$. In general, we want to define random orderings and we will need to take expectations with respect to S_k and S'_k, which will typically be random sets. The next subsection describes a class of processes for which the autocorrelation function can be expressed in terms of deterministic integrals. In certain cases analytic expressions can even be derived.

2.2 Orderings derived from a point process

In the sequel, we concentrate on a specific class of varying orderings that are defined by a driving point process Φ and a sequence of sets $U(\mathbf{x})$ for all values $\mathbf{x} \in D$. $U(\mathbf{x})$ defines the region in which points are relevant for determining the ordering at $\mathbf{x}$. The ordering, $\boldsymbol{\pi}(\mathbf{x})$, satisfies the condition

$$\|\mathbf{x} - \mathbf{z}_{\pi_1(\mathbf{x})}\| < \|\mathbf{x} - \mathbf{z}_{\pi_2(\mathbf{x})}\| < \|\mathbf{x} - \mathbf{z}_{\pi_3(\mathbf{x})}\| < \ldots,$$

where $\|\cdot\|$ is a distance measure and $\mathbf{z}_{\pi_i(\mathbf{x})} \in \Phi \cap U(\mathbf{x})$. We assume there are no ties, which is a.s. the case for *e.g.* Poisson point processes. Associating each atom $(V_i, \boldsymbol{\theta}_i)$ with the element of the point process $\mathbf{z}_i$ defines a marked point process from which we can define the distribution $F_{\mathbf{x}}$ for any $\mathbf{x} \in D$.

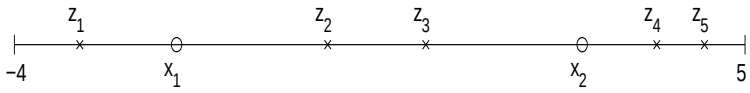

Figure 1: A section of a point process and two covariate values x_1 and x_2

Figure 1 illustrates this idea for a realisation of the point process, z, defined on the region (-4,5). If $U(x) = [-4, 5]$, the ordering at x_1 would be $1, 2, 3, 4, 5$ and the ordering at x_2 would be $4, 5, 3, 2, 1$. This choice of $U(x)$ leads to each ordering being a permutation. However, if $U(x) = [-4, x]$, the orderings at x_1 and x_2 would be 1 and $3, 2, 1$ respectively.

Specifying the type of point process allows us to derive more operational expressions for the autocorrelation function on the basis of (3).

The autocorrelation function now involves an expectation over the point process Φ. We make a slight change of notation by thinking of sets of points rather than indices so that now

$$T(\mathbf{x}_1, \mathbf{x}_2) = \Phi \cap U(\mathbf{x}_1) \cap U(\mathbf{x}_2)$$

and $A_{lk} = A_l(\mathbf{z}_k)$ which in general can be expressed as

$$A_l(\mathbf{z}) = \{\mathbf{w} \in \Phi \cap U(\mathbf{x}_l) | \|\mathbf{w} - \mathbf{x}_l\| < \|\mathbf{z} - \mathbf{x}_l\|\} \text{ for } \mathbf{z} \in \Phi \cap U(\mathbf{x}_l).$$

Similarly, define $S(\mathbf{z}) = A_1(\mathbf{z}) \cap A_2(\mathbf{z})$ and $S'(\mathbf{z}) = A_1(\mathbf{z}) \cup A_2(\mathbf{z}) - S(\mathbf{z})$ for $\mathbf{z} \in T(\mathbf{x}_1, \mathbf{x}_2)$. Thus, $S(\mathbf{z}) = \{\mathbf{w} \in T(\mathbf{x}_1, \mathbf{x}_2) | \|\mathbf{w} - \mathbf{x}_1\| < \|\mathbf{z} - \mathbf{x}_1\| \text{ and } \|\mathbf{w} - \mathbf{x}_2\| < \|\mathbf{z} - \mathbf{x}_2\|\}$, which clearly highlights that $S(\mathbf{z})$ groups all relevant points closer to $\mathbf{x}_1$ and $\mathbf{x}_2$ than to $\mathbf{z}$. These points are all associated with atoms

that precede the atom corresponding to $\mathbf{z}$ in the orderings at both $\mathbf{x}_1$ and $\mathbf{x}_2$. $S'(\mathbf{z})$ is its complement in the set of relevant points. The autocorrelation function is thus expressed as

$$\mathrm{Corr}(F_{\mathbf{x}_1}, F_{\mathbf{x}_2}) = \frac{2}{M+2} \mathrm{E}_\Phi \left[\sum_{z \in T(\mathbf{x}_1, \mathbf{x}_2)} \left(\frac{M}{M+2} \right)^{\#S(\mathbf{z})} \left(\frac{M}{M+1} \right)^{\#S'(\mathbf{z})} \right].$$

When Φ is a stationary point process, the refined Campbell theorem (Stoyan *et al.* 1995, p. 120) allows the autocorrelation to be expressed in terms of the Palm distribution of the point process (*e.g.* Stoyan *et al.* 1995, Ch. 7). Firstly, note that $\#S(\mathbf{z}) = \Phi(S(\mathbf{z}))$ and similarly for $S'(\mathbf{z})$. For a stationary point process with intensity λ, the refined Campbell theorem states that

$$\mathrm{E}_\Phi \left[\sum_{\mathbf{z} \in T(\mathbf{x}_1, \mathbf{x}_2)} f(\mathbf{z}, \Phi) \right] = \lambda \int_{U(\mathbf{x}_1) \cap U(\mathbf{x}_2)} \int f(\mathbf{z}, \varphi_{-\mathbf{z}}) \, P_o(d\varphi) \, d\mathbf{z},$$

where $P_o(d\varphi)$ is the Palm distribution of Φ at the origin and $\varphi_{-\mathbf{z}}$ is the realisation of Φ translated by $-\mathbf{z}$. In our case

$$f(\mathbf{z}, \varphi_{-\mathbf{z}}) = \left(\frac{M}{M+2} \right)^{\varphi_{-\mathbf{z}}(S_{-\mathbf{z}}(\mathbf{z}))} \left(\frac{M}{M+1} \right)^{\varphi_{-\mathbf{z}}(S'_{-\mathbf{z}}(\mathbf{z}))},$$

where $S_{-\mathbf{z}}(\mathbf{z})$ and $S'_{-\mathbf{z}}(\mathbf{z})$ are both translated by $-\mathbf{z}$, which leads to

$$\mathrm{Corr}(F_{\mathbf{x}_1}, F_{\mathbf{x}_2}) = \frac{2\lambda}{M+2} \int_{U(\mathbf{x}_1) \cap U(\mathbf{x}_2)} \int \left(\frac{M}{M+2} \right)^{\varphi_{-\mathbf{z}}(S_{-\mathbf{z}}(\mathbf{z}))} \left(\frac{M}{M+1} \right)^{\varphi_{-\mathbf{z}}(S'_{-\mathbf{z}}(\mathbf{z}))} P_o(d\varphi) \, d\mathbf{z}.$$

The simplest choice for the driving point process is the stationary Poisson process. In the sequel we show that this leads to a simpler form for the autocorrelation function that can be expressed in terms of deterministic integrals. These results are also useful when dealing with more general driving processes, such as Cox processes, as explained in Subsection 2.3.

Theorem 2 *If Φ follows a Poisson process with intensity λ, the autocorrelation can be expressed as*

$$\mathrm{Corr}(F_{\mathbf{x}_1}, F_{\mathbf{x}_2}) = \frac{2\lambda}{M+2} \int_{U(\mathbf{x}_1) \cap U(\mathbf{x}_2)} \exp\left\{ -\frac{\lambda}{M+1} d_{12}(\mathbf{z}) \right\} d\mathbf{z},$$

with $d_{12}(\mathbf{z}) = \nu(\{A_1(\mathbf{z})\}_{-\mathbf{z}}) + \nu(\{A_2(\mathbf{z})\}_{-\mathbf{z}}) - \frac{2}{M+2}\nu(S_{-\mathbf{z}}(\mathbf{z}))$, where $\{A_l(\mathbf{z})\}_{-\mathbf{z}}$ indicates the set $A_l(\mathbf{z})$ translated by $-\mathbf{z}$ and $\nu(\cdot)$ is the Lebesgue measure in d dimensions.

The autocorrelation function has been expressed in terms of an integral over a function of the areas of geometric objects, A_1, A_2 and S, which should help with its calculation. The following subsections describe two possible constructions which are useful in practical applications and for which an analytic expression for the autocorrelation function is available.

2.2.1 Permutations

A construction suitable for general smoothing problems and spatial modelling is obtained through defining $D \subset \mathbb{R}^d$ ($d = 2$ for most spatial problems) and $U(\mathbf{x}) = D$ for all values of $\mathbf{x}$. In one dimension ($d = 1$), we can derive an analytic form for the autocorrelation function.

Corollary 1 *Let Φ be Poisson with intensity λ, $D \subset \mathbb{R}$ and $U(x) = D$ for all x. Then we obtain*

$$Corr(F_{x_1}, F_{x_2}) = \left(1 + \frac{2\lambda h}{M+2}\right) \exp\left\{\frac{-2\lambda h}{M+1}\right\},$$

where $h = |x_1 - x_2|$ is the distance between x_1 and x_2.

Note the unusual form of the correlation structure above. It is decreasing in the distance, but is the weighted sum of a Matérn correlation function with smoothness parameter $3/2$ (with weight $(M+1)/(M+2)$) and an exponential correlation function (with weight $1/(M+2)$), which is a less smooth member of the Matérn class, with smoothness parameter $1/2$. So for $M \to 0$ the correlation function will tend to the arithmetic average of both and for large M the correlation structure will behave like a Matérn with smoothness parameter $3/2$.

In higher dimensions, for $d \geq 2$, the autocorrelation function can be expressed as a two-dimensional integral, as detailed in Appendix B.

2.2.2 Arrivals ordering

A framework which might be considered more suitable for modelling time series is obtained by choosing $D = \mathbb{R}$ and $U(x) = (-\infty, x]$. In this case only those points with arrival times before x will be used in determining the ordering at time x.

Corollary 2 *Let Φ be Poisson with intensity λ, $D \subset \mathbb{R}$ and $U(x) = (-\infty, x]$ for all x. Then we obtain*

$$Corr(F_{x_1}, F_{x_2}) = \exp\left\{-\frac{\lambda h}{M+1}\right\},$$

where h is as defined in Corollary 1.

Thus, this construction leads to the well-known exponential correlation structure.

The relative ordering of the points that are already in the representation remains the same as time goes on. At each arrival a new point is added, which will be allocated the first rank in the ordering, with weight $p_1 = V_{\pi_1(x_2)}$. Thus, if x_1 is the previous arrival time and the new arrival time corresponding to atom $\boldsymbol{\theta}_{\pi_1(x_2)}$ is $x_2 > x_1$, then

$$F_{x_2} \stackrel{d}{=} \left(1 - V_{\pi_1(x_2)}\right) F_{x_1} + V_{\pi_1(x_2)} \delta_{\boldsymbol{\theta}_{\pi_1(x_2)}}.$$

This form is reminiscent of a first-order random coefficient autoregressive process with jumps.

Throughout this Subsection 2.2, the correlation depends on λ and M roughly through the ratio $\lambda/(M+1)$. This is not surprising: for small M, only the first few atoms will matter and then we only need a few points per unit volume to induce a certain correlation. If M is larger, we need to re-order many atoms to change the distribution appreciably, and thus we need a large λ to obtain the same correlation.

2.3 More flexible autocorrelation functions

An attractive option for defining more general forms of autocorrelation function is to use a Cox process as the driving point process Φ. Examples include mixed Poisson processes and Poisson cluster processes. Møller (2003) defines shot noise Cox processes which could generate a very wide class of potential forms for the autocorrelation function. We assume that Φ follows a Poisson point process conditional on the intensity Λ, which is a random measure drawn from a distribution Q. For example, a mixed Poisson process arises if Q has a discrete distribution with a finite expectation. Stationarity of Φ will follow from the stationarity of Λ. Standard results are readily available for the Palm distribution of a Cox process, which is

$$\lambda P_o(Y) = \int \mu P_o^{\mu}(Y) Q(d\mu)$$

where P_o^{μ} is the Palm distribution of a Poisson process with intensity μ and $\lambda = \int \mu Q(d\mu)$.

The dependence structure is now characterized by

$$\begin{aligned}\text{Corr}(F_{\mathbf{x}_1}, F_{\mathbf{x}_2}) &= \frac{2\lambda}{M+2} \int_{U(\mathbf{x}_1)\cap U(\mathbf{x}_2)} \int \left(\frac{M}{M+2}\right)^{\varphi_{-\mathbf{z}}(S_{-\mathbf{z}}(\mathbf{z}))} \left(\frac{M}{M+1}\right)^{\varphi_{-\mathbf{z}}(S'_{-\mathbf{z}}(\mathbf{z}))} P_o(d\varphi)\, d\mathbf{z} \\ &= \frac{2}{M+2} \int \mu \int_{U(\mathbf{x}_1)\cap U(\mathbf{x}_2)} \exp\left\{-\frac{\mu}{M+1} d_{12}(\mathbf{z})\right\} Q(d\mu)\, d\mathbf{z}.\end{aligned}$$

With the arrivals construction, for example, this correlation function simplifies to

$$\text{Corr}(F_x, F_{x+h}) = \frac{2}{M+2} \int \frac{\mu}{\lambda} \exp\left\{-\frac{\mu h}{M+1}\right\} Q(d\mu).$$

3 Mixtures of Order-based Dependent Dirichlet processes

The Dirichlet process provides random distributions with discrete realisations. The mixture of Dirichlet process model (Antoniak, 1974) provides an alternative framework which can generate absolutely continuous distributions. This model has proved popular in applied Bayesian nonparametric work. It can be expressed hierarchically for observation i as

$$\begin{aligned} p(y_i|\boldsymbol{\psi}_i) &= f(y_i|\boldsymbol{\psi}_i) \\ \boldsymbol{\psi}_i &\overset{i.i.d.}{\sim} F \\ F &\sim \text{DP}(MH). \end{aligned}$$

The πDDP can be used to extend this model to spatial, time series or regression problems by simply replacing F by $F_{\mathbf{x}_i}$, given by

$$F_{\mathbf{x}} \sim \pi\text{DDP}(MH, \lambda),$$

where the notation $\pi\text{DDP}(MH, \lambda)$ denotes a πDDP characterised by mass parameter M, centring distribution H and an ordering induced by a Poisson point process with intensity λ.

This model includes the Bayesian Partition Model (see *e.g.* Denison *et al.* 2002) as a limiting case. As $M \to 0$, the random distribution tends to a Dirac measure at the first element of the ordering.

Observations whose covariate values are closest to a particular point will have equal values of ψ_i. The same model would arise by defining a Voronoi tessellation of the domain using the points as centres and assuming that all observations with covariates in the same region have common parameter values. This model was proposed for general non-linear regression problems and has been used for spatial mapping problems (*e.g.* Ferreira *et al.* 2002 and Knorr-Held and Raßer 2000). As the intensity $\lambda \to 0$, we will not get any switches in the ordering and $F_{\mathbf{x}}$ will no longer depend on $\mathbf{x}$. Thus, we will recover the mixture of Dirichlet process model.

4 Prior distributions for M and λ

In general, inference about the parameters M and λ is not possible when we directly model continuous observations through a πDDP. However, in the mixture of Dirichlet processes model inference is possible and M can be interpreted as controlling the probability that $\psi_i = \psi_j$ for $i \neq j$. Consequently, we will make inference in the model described above. The prior distribution for M is an inverted Beta distribution

$$p(M) = \frac{n_0^\eta \Gamma(2\eta)}{\Gamma(\eta)^2} \frac{M^{\eta-1}}{(M+n_0)^{2\eta}},$$

which was introduced by Griffin and Steel (2004) and where the hyperparameter $n_0 > 0$ is the prior median of M and the prior variance of M (which exists if $\eta > 2$) is a decreasing function of η. It implies that $M/(M+n_0)$ follows a $\mathrm{Beta}(\eta, \eta)$ distribution and that $\phi = 1/(M+1)$ has a Gauss hypergeometric distribution (see Johnson *et al.* 1995, p. 253):

$$p(\phi) = \frac{n_0^\eta \Gamma(2\eta)}{\Gamma(\eta)^2} (1-\phi)^{\eta-1} \phi^{\eta-1} (1+(n_0-1)\phi)^{-2\eta}.$$

The parameter $\phi \in (0,1)$, which appears in equation (2), is of interest as it relates the variance of the measure F to the variance of the measure H (our parametric centring distribution) and can be interpreted as a measure of the appropriateness of the parametric model H. Values of ϕ away from zero indicate a failure of the model H to capture the conditional (with respect to $\mathbf{x}$) distribution of the data at hand.

An independent prior distribution on the autocorrelation function specifies a corresponding prior distribution for λ given M. The prior distribution can be defined for any valid stationary autocorrelation function and is specified by choosing a value, $t^\star = \|\mathbf{x}_1 - \mathbf{x}_2\|$, for which the correlation $\mathrm{Corr}(F_{\mathbf{x}_1}, F_{\mathbf{x}_2})$ follows a uniform prior distribution. In the case of the arrivals construction, this choice implies that $\lambda \sim \mathrm{Exp}(t^\star/(M+1))$ and that the correlation at distance h is distributed as $\mathrm{Beta}(t^\star/h, 1)$. For the permutation construction with $d = 1$, the induced distribution of λ is

$$p(\lambda) = \frac{2t^\star(2t^\star\lambda + 1)}{(M+1)(M+2)} \exp\left\{-\frac{2t^\star}{M+1}\lambda\right\}.$$

If $d > 1$, the autocorrelation function is not available in closed form and so there will be no closed form expression for the implied prior on λ. In that case, we will approximate this prior numerically.

5 Computational method

We assume that we have observed values for $\mathbf{y} = (y_1, \ldots, y_n)$, associated with covariate values $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$. Early work on MCMC-based inference for Dirichlet process mixture models is described in MacEachern (1998) and make use of a Polya urn representation. In this case, no truncation is required for posterior and predictive inference. Gelfand and Kottas (2002) discuss how inference can then be conducted on general functionals of the random distribution. The methods described in this paper follow more closely the truncation-based method described in Ishwaran and James (2001) and many ideas carry over directly. The idea of truncating the Sethuraman representation for simulation purposes was proposed in Muliere and Tardella (1998). An added complication in the method for our model is the need to sample the point process $\mathbf{z}$ and the intensity parameter λ. For simulation purposes, we truncate the Poisson point process to $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_T)$ (further details are given later in this section). In general, we have parameters $\mathbf{z}, \lambda, \mathbf{s}, \boldsymbol{\theta}, \mathrm{V}, M$ where $\mathbf{s}$ is an n-dimensional vector for which $\boldsymbol{\psi}_i = \boldsymbol{\theta}_{s_i}$. Additionally, the distribution H or the density f may have parameters that we wish to conduct inference on. In contrast to Ishwaran and James (2001), we will use the Gibbs sampler for the posterior distribution marginalised over the parameters V and, where possible, over the parameters $\boldsymbol{\theta}$. Models where the second marginalisation is possible are typically called conjugate Dirichlet process mixture models. For non-conjugate models, the parameter vector $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T$ must also be simulated. The updating of these parameters follows from standard Dirichlet process method described by Ishwaran and James (2001). The effects on other parts of the sampler will be discussed in the sequel.

A feature of the method described in Ishawaran and James (2001) is the need to truncate the stick-breaking representation at an appropriate value N (recently, Papaspiliopoulos and Roberts (2004) have developed an algorithm where truncation is not necessary). Since the weights of the discrete distribution $F_{\mathbf{x}}$ are stochastically ordered, Ishwaran and Zarepour (2000) suggest choosing a value of N that bounds the expectation of the error $\sum_{i=N+1}^{\infty} p_i$, which has the form $(M/(M+1))^N$. In our case, it is more natural to define a truncated region (which we will call the computational region) for the point process $\mathbf{z}$ that includes the range of the covariates $\mathbf{x}$. The truncation error will be largest at the extreme values of this region. Let us first assume that x is one-dimensional, that the smallest and largest x values are d_a and d_b, respectively, that we choose the computational region (a, b) and that z follows a Poisson process with intensity λ. The expectation of the error $\sum_{i=N+1}^{\infty} p_i$ at $x = d_b$ will then be $\exp\{-\lambda(b - d_b)/(M+1)\}$. If we want fix the error at, say, $\epsilon \in (0, 1)$ then we need to choose $b = d_b - \{(M+1)\log\epsilon\}/\lambda$ and similarly $a = d_a + \{(M+1)\log\epsilon\}/\lambda$. This choice of truncation leads to the nice property that the number of points in the computational region outside the data region of x is independent of λ which avoids some overconditioning issues.

If $d > 1$, we choose a bounding box say $(a_1, b_1) \times (a_2, b_2) \times \cdots \times (a_d, b_d)$ as the computational region and let d_{ai} and d_{bi} respectively be the minimum and maximum values of x in dimension i. The truncation error will be greatest at the corners of the box. If we define $a_i = d_{ai} - r$ and $b_i = d_{bi} + r$, the truncation error ϵ will be $\exp\{-\frac{\lambda}{M+1}\frac{2\pi^{d/2}}{\Gamma(d/2)d}\left(\frac{r}{2}\right)^d\}$, which implies a value of $r = 2\left(\frac{\Gamma(d/2)d}{2\pi^{d/2}}\frac{M+1}{\lambda}\log\frac{1}{\epsilon}\right)^{1/d}$.

At this point, it is useful to define some notation. For a subset C of $\mathcal{I} = \{1, \ldots, n\}$, define the summaries $n_l(C)$ to be the number of i's in C for which $s_i = l$ (*i.e.* the number of points in the

subset allocated to a point $\mathbf{z}_l$) and $W_l(C) = \#\{i \in C$ such that there exists $k < j$ for which $\pi_k(\mathbf{x}_i) = l$ where $\pi_j(\mathbf{x}_i) = s_i\}$ (*i.e.* the number of observations for which l appears before s_i in the ordering at $\mathbf{x}_i$). For $\mathbf{w} = (w_1, \ldots, w_k)$ define $\mathbf{w}_{-i} = (w_1, \ldots, w_{i-1}, w_{i+1}, \ldots, w_k)$. Finally, the posterior distribution is

$$p(\mathbf{s}, \mathbf{z}, M, \lambda | \mathbf{y}, \mathbf{x}) \propto p(\mathbf{y}|\mathbf{s}) p(\mathbf{s}|M, \mathbf{z}, \mathbf{x}) p(\mathbf{z}|\lambda) p(\lambda) p(M),$$

where

$$p(\mathbf{y}|\mathbf{s}) = \prod_{l=1}^{T} \left[\int \prod_{\{1 \leq i \leq n | s_i = l\}} f(y_i|\boldsymbol{\psi})\, dH(\boldsymbol{\psi}) \right]$$

and

$$p(\mathbf{s}|M, \mathbf{z}, \mathbf{x}) = M^T \prod_{l=1}^{T} \frac{\Gamma(n_l(\mathcal{I}) + 1)\Gamma(W_l(\mathcal{I}) + M)}{\Gamma(n_l(\mathcal{I}) + 1 + W_l(\mathcal{I}) + M)}.$$

5.1 Updating s

The full conditional distribution for $\mathbf{s}$ is a discrete distribution and can be simulated directly. To simplify notation, we define $\boldsymbol{\zeta} = (M, \mathbf{z}, \mathbf{x})$. The probabilities are

$$P(s_i = l \,|\mathbf{s}_{-i}, \mathbf{y}, \boldsymbol{\zeta}) = 0$$

if $\mathbf{z}_l \notin U(\mathbf{x}_i)$ and otherwise

$$\begin{aligned} P(s_i = l \,|\mathbf{s}_{-i}, \mathbf{y}, \boldsymbol{\zeta}) &\propto p(y_i \,|s_i = l, \mathbf{s}_{-i}, \mathbf{y}_{-i})\, P(s_i = l \,|\mathbf{s}_{-i}, \boldsymbol{\zeta}) \\ &= \frac{\int f(y_i|\boldsymbol{\psi}) \prod_{\{j \neq i | s_j = l\}} f(y_j|\boldsymbol{\psi})\, dH(\boldsymbol{\psi})}{\int \prod_{\{j \neq i | s_j = l\}} f(y_j|\boldsymbol{\psi})\, dH(\boldsymbol{\psi})} \frac{n_l(\mathcal{I}_{-i}) + 1}{M + W_l(\mathcal{I}_{-i}) + n_l(\mathcal{I}_{-i}) + 1} \\ &\quad \times \prod_{j < m(l)} \frac{M + W_{\pi_j(\mathbf{x})}(\mathcal{I}_{-i})}{M + W_{\pi_j(\mathbf{x})}(\mathcal{I}_{-i}) + n_{\pi_j(\mathbf{x})}(\mathcal{I}_{-i}) + 1} \end{aligned}$$

where $\pi_{m(l)}(\mathbf{x}) = l$.

We now turn our attention to updating the point process $\mathbf{z}$, the intensity λ and the mass parameter M. The point process $\mathbf{z}$ is updated using a hybrid Reversible Jump step (Green, 1995). Slighly unusually, we also suggest updating a subset of the allocation variables $\mathbf{s}$ jointly with $\mathbf{z}$. The two parameters, λ and M, could be updated using a standard Metropolis-within-Gibbs methods but we suggest also updating the point process (and the allocation variables). In both cases, these more complicated methods should avoid slow mixing of the chain. In all cases a parameter with a dash will represent the proposed value of that parameter or summary.

5.2 Updating z

There are three possible updates in the Reversible Jump MCMC sampler: move a current point, birth of a new point or death of a current point. For the last two proposals, the method assumes that the locations, $\boldsymbol{\theta}$, can be marginalised from the posterior distribution. Extensions for non-conjugate problems are discussed at the end of 5.2.2.

5.2.1 Move

A point $\mathbf{z}_l$ is chosen at random and updated using a random walk Metropolis-Hastings move by adding a normally distributed random variable with mean zero and a tuning covariance matrix to $\mathbf{z}_l$. The update is rejected if the point moves outside the computational region or if $\mathbf{z}_l \notin U(\mathbf{x}_i)$ for i such that $s_i = l$. Otherwise, the acceptance probability is

$$\min\left\{1, \prod_{i=1}^{T} \frac{n_i(\mathcal{I}) + 1 + W_i(\mathcal{I}) + M}{n_i(\mathcal{I}) + 1 + W_i'(\mathcal{I}) + M}\right\}.$$

5.2.2 Birth and death

The birth and death moves come as a pair that maintain reversibility of the sampler. After a point has been added (birth) or removed (death) from the point process the allocations of certain observations are updated. For the death move, a point, $\mathbf{z}_j$, is chosen uniformly from $\mathbf{z}_1, \ldots, \mathbf{z}_T$. To complete the move, the observations allocated to $\mathbf{z}_j$ must be re-allocated. The set of possible points is restricted to be close to $\mathbf{z}_j$ and is defined by $\mathcal{T}_D = \{i | \|\mathbf{z}_i - \mathbf{z}_j\| \leq c, i \neq j\}$. The observations that need to be re-allocated are $\mathcal{I}_D = \{i | s_i = j\} = \{i_1, \ldots, i_{n_j}\}$. We will work sequentially through this set and re-allocate according to the discrete distribution with probabilities proportional to

$$P\left(s'_{i_k} = l \,\middle|\, \mathcal{Y}^{(k)}, \mathcal{S}^{(k)}, \boldsymbol{\zeta}\right) = 0$$

if $\mathbf{z}_l \notin U(\mathbf{x}_{i_k})$ and otherwise

$$\begin{aligned} P\left(s'_{i_k} = l \,\middle|\, \mathcal{Y}^{(k)}, \mathcal{S}^{(k)}, \boldsymbol{\zeta}\right) &\propto p\left(y_{i_k} \,\middle|\, \mathcal{Y}^{(k-1)}, s'_{i_k} = l, \mathcal{S}^{(k)}\right) P\left(s_{i_k} = l \,\middle|\, \mathcal{S}^{(k)}, \boldsymbol{\zeta}\right) \\ &= \frac{\int f(y_{i_k}|\boldsymbol{\psi}) \prod_{\{i \in \mathcal{I}^{(k)} | s'_i = l\}} f(y_i|\boldsymbol{\psi})\, dH(\boldsymbol{\psi})}{\int \prod_{\{i \in \mathcal{I}^{(k)} | s'_i = l\}} f(y_i|\boldsymbol{\psi})\, dH(\boldsymbol{\psi})} \frac{n_l(\mathcal{I}^{(k)}) + 1}{M + W_l(\mathcal{I}^{(k)}) + n_l(\mathcal{I}^{(k)}) + 1} \\ &\quad \times \prod_{j < m(l)} \frac{M + W_{\pi_j(\mathbf{x})}(\mathcal{I}^{(k)})}{M + W_{\pi_j(\mathbf{x})}(\mathcal{I}^{(k)}) + n_{\pi_j(\mathbf{x})}(\mathcal{I}^{(k)}) + 1}, \qquad l \in \mathcal{T}_D \end{aligned} \tag{4}$$

where $\pi_{m(l)}(\mathbf{x}) = l$ and $\mathcal{I}^{(k)} = (\mathcal{I} - \mathcal{I}_D) \cup \{i_1, \ldots, i_{k-1}\}$, $\mathcal{Y}^{(k)} = \{y_i | s_i \neq j\} \cup \{y_{i_1}, \ldots, y_{i_k}\}$ and $\mathcal{S}^{(k)} = \{s_i | s_i \neq j\} \cup \{s'_{i_1}, \ldots, s'_{i_{k-1}}\}$. Without requiring additional user input, this provides an efficient solution to the problem of Gibbs steps, which have a tendency to get stuck in local modes. See Dahl (2003) for a discussion of a similar idea, and alternatives, when sampling a Dirichlet process mixture model.

In the case of the reverse birth move, a new point $\mathbf{z}_{T+1}$, is chosen uniformly over the computational region. Reversibility suggests that the observations that could be re-allocated are the ones which are allocated to points in the set $\mathcal{T}_B = \{i | \|\mathbf{z}_i - \mathbf{z}_{T+1}\| \leq c, i = 1, \ldots, T\}$. If this set is empty then the proposal is rejected. Let $\mathcal{I}_B = \{i | s_i \in \mathcal{T}_B\} = \{i_1, \ldots, i_m\}$ be the points that can be re-allocated. Then the elements of $\mathcal{I}_B$ are allocated sequentially. The observation i_k is allocated to $\mathbf{z}_{T+1}$ with probability proportional to

$$P\left(s'_{i_k} = T + 1 \,\middle|\, \mathcal{Y}_B^{(k)}, \mathcal{S}_B^{(k)}, \boldsymbol{\zeta}\right)$$

and not re-allocated with probability proportional to

$$\sum_{j\in\mathcal{T}_B} P\left(s'_{i_k}=j\,\middle|\,\mathcal{Y}_B^{(k)},\mathcal{S}_B^{(k)},\boldsymbol{\zeta}\right),$$

where $\mathcal{S}_B^{(k)}=\{s_i|i\notin\mathcal{I}_B\}\cup\{s'_{i_1},\ldots,s'_{i_{k-1}}\}$ and $\mathcal{Y}_B^{(k)}=\{y_i|i\notin\mathcal{I}_B\}\cup\{y_{i_1},\ldots,y_{i_k}\},k=1,\ldots,m$. The acceptance rate for the birth move can be calculated using the following argument. Let the proposed new point be $\mathbf{z}_{T+1}$. The probability of the birth proposal can be written as

$$q(\mathbf{s},\mathbf{s}')=\prod_{k=1}^{m}\frac{\left(\sum_{j\in\mathcal{T}_B}P\left(s'_{i_k}=j|\mathcal{Y}_B^{(k)},\mathcal{S}_B^{(k)},\boldsymbol{\zeta}\right)\right)^{\mathrm{I}(s'_{i_k}=s_{i_k})}P\left(s'_{i_k}=T+1|\mathcal{Y}_B^{(k)},\mathcal{S}_B^{(k)},\boldsymbol{\zeta}\right)^{\mathrm{I}(s'_{i_k}=T+1)}}{P\left(s'_{i_k}=T+1|\mathcal{Y}_B^{(k)},\mathcal{S}_B^{(k)},\boldsymbol{\zeta}\right)+\sum_{j\in\mathcal{T}_B}P\left(s'_{i_k}=j|\mathcal{Y}_B^{(k)},\mathcal{S}_B^{(k)},\boldsymbol{\zeta}\right)}$$

where I denotes the indicator function and the probability of the reverse proposal can be written as

$$q(\mathbf{s}',\mathbf{s})=\prod_{k=1}^{m}\left(\frac{p\left(s_{i_k}|\mathcal{Y}_B^{(k)},\mathcal{S}_B^{(k)},\boldsymbol{\zeta}\right)}{\sum_{j\in\mathcal{T}_B}P\left(s_{i_k}=j|\mathcal{Y}_B^{(k)},\mathcal{S}_B^{(k)},\boldsymbol{\zeta}\right)}\right)^{\mathrm{I}(s'_{i_k}=T+1)}.$$

This leads to the acceptance probability

$$\min\left\{1,\frac{p(\mathbf{y}|\mathbf{s}')p(\mathbf{s}'|M,\mathbf{z}',\mathbf{x})q(\mathbf{s}',\mathbf{s})}{p(\mathbf{y}|\mathbf{s})p(\mathbf{s}|M,\mathbf{z}',\mathbf{x})q(\mathbf{s},\mathbf{s}')}\right\}.$$

The acceptance rate for the death move can be calculated in a similar way.

The method above constructs a proposal distribution for $\mathbf{z}$ and $\mathbf{s}$. In the non-conjugate case, we need a proposal for $\mathbf{z}$, $\mathbf{s}$ and $\boldsymbol{\theta}$. A simple method for the birth move, proposes $\boldsymbol{\theta}'_i=\boldsymbol{\theta}_i$ for $1\leq i\leq T$ and proposes the new value $\boldsymbol{\theta}'_{T+1}$ from the centring distribution H. Then $\mathbf{s}'$ could be proposed using a modified version of (4) where $p(y_{i_k}|\mathcal{Y}^{(k-1)},s'_{i_k}=l,\mathcal{S}^{(k)})$ is replaced by $p(y_{i_k}|s'_{i_k}=l,\boldsymbol{\theta}'_l)$. This proposal leaves the acceptance probability unchanged but may not work well in some problems. In particular, conditioning on $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_T$ and drawing $\boldsymbol{\theta}_{T+1}$ may lead to little re-allocation in proposing $\mathbf{s}'$. An alternative method would propose $\mathbf{s}'$ before $\boldsymbol{\theta}'$ using an approximation to (4). Then $\boldsymbol{\theta}'$ could be proposed from an approximation of $p(\boldsymbol{\theta}'|\mathbf{s}',\mathbf{y},\mathbf{x})$. The success of either approach will depend greatly on the problem at hand.

5.3 Updating M

The definition of the computational region means that the number of points which are in the computational region but not in the data region depends on M. The usual Gibbs step would be affected by the current number of these points which has been chosen conditional on the current value of M. Since this definition is chosen to avoid edge-effects, it seems undesirable that it should also affect the sampler. The following update removes the associated term from the acceptance probability. A new value of M is proposed such that $\log M'\sim\mathrm{N}(\log M,\sigma_M^2)$ where σ_M^2 can be chosen to control the overall acceptance rate of this step. If $M'>M$ then the computational region is expanded and the unobserved part of the Poisson process is sampled. If $M'<M$, the natural reverse contracts the computational region and

removes from the sampler those points that now fall outside this region. If the latter points have any observations allocated to them, the proposal is rejected. This move is in effect a reversible jump move where we sample extra points from the prior distribution. The acceptance probability in this case is

$$\min\left\{1, \frac{M'}{M}\frac{p(M'|\lambda)}{p(M|\lambda)}\prod_{i=1}^{T}\frac{n_i(\mathcal{I})+1+W_i(\mathcal{I})+M}{n_i(\mathcal{I})+1+W_i'(\mathcal{I})+M'}\right\}.$$

5.4 Updating λ

The parameter λ can sometimes suffer from the problem of overconditioning (Papaspiliopoulos *et al.* 2003), which occurs because the full conditional for λ depends on $\mathbf{z}$ which itself is latent. The lack of direct data information for λ can lead to slow mixing chains. Separate sampling schemes for λ are described for $d = 1$ (*i.e.* univariate x) and $d > 1$. In both cases, we make use of the ideas described in Papaspiliopoulos *et al.* (2003) for sampling Poisson processes. Each point of the Poisson process $\mathbf{z}_i$ is given a mark t_i which is uniformly distributed on $(0, 1)$. A new value of the parameter $\log\lambda' \sim \mathrm{N}(\log\lambda, \sigma_\lambda^2)$ is proposed. For $d = 1$, if $\lambda' < \lambda$ those points in the data region for which $t_i > \lambda'/\lambda$ are removed from the point process, otherwise $t_i' = t_i\lambda/\lambda'$ and if $\lambda' > \lambda$ then a new Poisson process with intensity $\lambda' - \lambda$ is drawn in the data region. The value of t_i' for each new point $i = T+1, \ldots, T'$ is proposed from a uniform distribution on the region$[\lambda/\lambda', 1)$ and the proposed value for $i = 1, \ldots, T$ is $t_i' = t_i\lambda/\lambda'$. The proposed values for points outside the data region are as follows: if $z_i < d_a$, $z_i' = d_a + (z_i - d_a)\frac{d_a - a'}{d_a - a}$ and if $z_i > d_b$ then $z_i' = d_b + (z_i - d_b)\frac{b' - d_b}{b - d_b}$. If $\lambda' > \lambda$, the proposed points are worked through sequentially. For each new point, the allocations are updated as in the birth step introduced in Subsection 5.2.2. If $\lambda' < \lambda$, the allocations are updated for each deleted point in turn as in the death step.

If $d > 1$, the number of points outside the data region is not independent of λ. Consequently, the updating mechanism is the same for points in the data region but outside the data region a different scheme is used. If $\lambda' < \lambda$, all points outside the new computational region are deleted and all point inside the new computational region for which $t_i > \lambda'/\lambda$ are deleted, otherwise we assign $t_i' = t_i\lambda/\lambda'$. If $\lambda' > \lambda$, a new Poisson process with intensity $\lambda' - \lambda$ is drawn on the computational region defined by the previous parameter values and a Poisson process with intensity λ' is drawn on the part of the computational region that has been added. Once again, the proposed value t_i' for each new point $i = T + 1, \ldots, T'$ is from a uniform distribution on the region $[\lambda/\lambda', 1)$ and the proposed value for $i = 1, \ldots, T$ is $t_i' = t_i\lambda/\lambda'$.

For any value of d, the acceptance rate for $\lambda' > \lambda$ can be calculated in the following way. Let the points added to the process be $\mathbf{z}'_{T+1}, \ldots, \mathbf{z}'_{T'}$ and let $\mathbf{z}'_1, \ldots, \mathbf{z}'_T$ be the position of $\mathbf{z}_1, \ldots, \mathbf{z}_T$ after potential moves. Let $\mathcal{T}_j = \{i : \|\mathbf{z}_i' - \mathbf{z}'_{T+j}\| \leq c, i = 1, \ldots, T\}$. If this set is empty then the proposal is rejected. Let $\mathcal{I}_j = \{i : s_i \in \mathcal{T}_j\} = \{i_{j1}, \ldots, i_{jm_j}\}$ be the points that can be re-allocated. For $k = 1, \ldots, m_j$, let $\mathcal{S}_j^{(k)} = \{s_i : i \notin \mathcal{I}_j\} \cup \{s'_{i_{j1}}, \ldots, s'_{i_{j(k-1)}}\}$ and $\mathcal{Y}_j^{(k)} = \{y_i : i \notin \mathcal{I}_j\} \cup \{y_{i_{j1}}, \ldots, y_{i_{jk}}\}$. The probability of the birth proposal can be written as

$$q(\mathbf{s}, \mathbf{s}') = \prod_{l=1}^{T'-T}\prod_{k=1}^{m_l}\frac{\left(\sum_{j\in\mathcal{T}_l} P\left(s'_{i_{lk}} = j|\mathcal{Y}_l^{(k)}, \mathcal{S}_l^{(k)}, \boldsymbol{\zeta}\right)\right)^{\mathrm{I}(s'_{i_{lk}} = s_{i_{lk}})} P\left(s'_{i_{lk}} = T+1|\mathcal{Y}_l^{(k)}, \mathcal{S}_l^{(k)}, \boldsymbol{\zeta}\right)^{\mathrm{I}(s'_{i_{lk}} = T+1)}}{P\left(s'_{i_{lk}} = T+1|\mathcal{Y}_l^{(k)}, \mathcal{S}_l^{(k)}, \boldsymbol{\zeta}\right) + \sum_{j\in\mathcal{T}_l} P\left(s'_{i_{lk}} = j|\mathcal{Y}_l^{(k)}, \mathcal{S}_l^{(k)}, \boldsymbol{\zeta}\right)}$$

and the probability of the reverse proposal can be written as

$$q(\mathbf{s}', \mathbf{s}) = \prod_{l=1}^{T'-T} \prod_{k=1}^{m_l} \left(\frac{p\left(s_{i_{lk}} | \mathcal{Y}_l^{(k)}, \mathcal{S}_l^{(k)}, \boldsymbol{\zeta}\right)}{\sum_{j \in \mathcal{T}_l} P\left(s_{i_{lk}} = j | \mathcal{Y}_l^{(k)}, \mathcal{S}_l^{(k)}, \boldsymbol{\zeta}\right)} \right)^{\mathrm{I}(s'_{i_{lk}} = T+1)},$$

leading to the acceptance rate

$$\frac{p(\mathbf{y}|\mathbf{s}')p(\mathbf{s}'|M, \mathbf{z}', \mathbf{x})q(\mathbf{s}', \mathbf{s})\lambda' p(M|\lambda')p(\lambda')}{p(\mathbf{y}|\mathbf{s})p(\mathbf{s}|M, \mathbf{z}, \mathbf{x})q(\mathbf{s}, \mathbf{s}')\lambda p(M|\lambda)p(\lambda)}.$$

6 Applications

Here we describe three rather different settings where mixtures of order-based dependent Dirichlet processes prove useful. We use generated data from a regression example with a scalar covariate, observed time series data where we allow for volatility changing over time, and spatial temperature data.

Throughout, we use a Poisson point process with intensity λ to generate the ordering in combination with the permutations construction for the regression and spatial applications and the arrivals construction for the time series application.

6.1 Regression Modelling

A model for curve-fitting can be defined by extending the model for density estimation described by Escobar and West (1995). They use a Dirichlet process mixture of normals which can be extended simply by defining an order-based DDP in place of the Dirichlet process. In contrast to their work, we will assume a common variance for the conditional distribution of the observations y_i. The model can be expressed as the following hierarchical model

$$\begin{aligned} y_i &\sim \mathrm{N}(\mu_i, \sigma^2) \\ \mu_i &\sim F_{\mathbf{x}_i} \\ F_{\mathbf{x}} &\sim \pi\mathrm{DDP}(MH, \lambda). \end{aligned} \tag{5}$$

The model is centred in the usual sense since $F_{\mathbf{x}}$ follows a Dirichlet process for any $\mathbf{x}$ and so marginalising over F gives

$$p(y_i|\mathbf{x}_i) = \int \mathrm{N}(\mu_i, \sigma^2) dH(\mu_i).$$

This model limits to a piecewise constant model as $M \to 0$. An alternative model can be defined by also modelling σ with the Dirichlet process.

The following simulated example illustrates the flexibility of the Dirichlet process to adapt to the properties of the phenomenon under consideration. A sample of $n = 100$ data-points was generated randomly around a sine curve in the interval $D = [0, 1]$ from

$$p(y_i|x_i) = \mathrm{N}(y_i| \sin(2\pi x_i), 0.01).$$

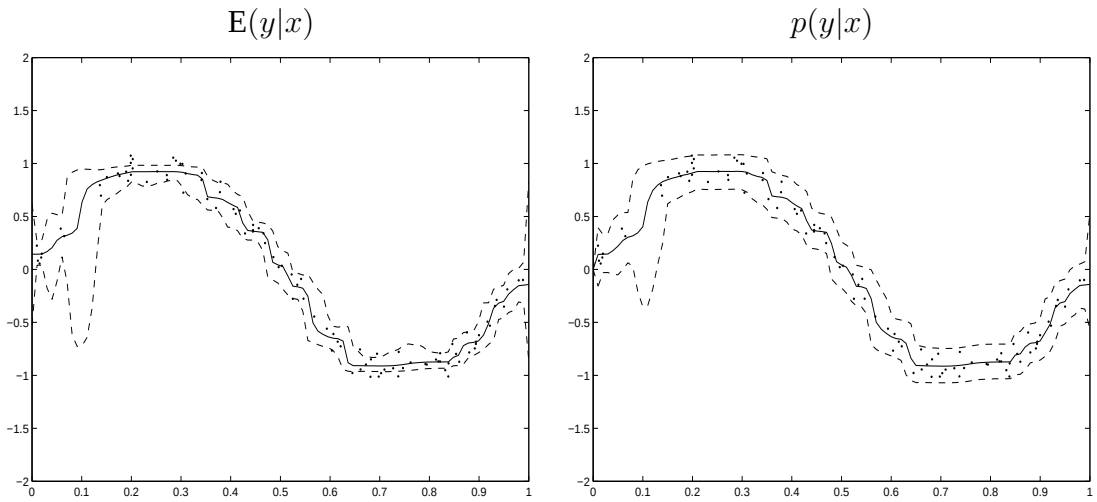

Figure 2: Sine curve regression data: The posterior distribution of $\mathrm{E}(y|x)$ and the predictive distribution $p(y|x)$ summarised by the median (solid line) and the 95% credible intervals (dashed lines) as a function of x. The $n = 100$ data points are indicated as dots. We have chosen $t^\star = 0.2$ in the prior for λ.

We fit these data (indicated by dots in Figure 2) with the model in (5). For the centring distribution H we take $\mathrm{N}(0, \sigma^2/\kappa)$ where $\kappa \sim \mathrm{IG}(0.001, 0.00001)$ ($\mathrm{IG}(\alpha, \beta)$ denotes an inverse gamma distribution with shape parameter α and scale β) and the prior distribution on σ is $\mathrm{IG}(0.001, 0.001)$. We take the values $n_0 = 1$ and $\eta = 0.5$ in the prior for M. We use the permutations construction to induce the ordering to vary with x and experiment with various values of $t^\star$ in the prior on λ.

The estimate of the function is illustrated in Figure 2 which presents the posterior median and 95% credible region of $\mathrm{E}[y|x]$, as well as the predictive median and credible region. The results illustrate the ability of the dependent Dirichlet process to fit the data under consideration despite its simple form.

Posterior distributions on σ and other quantities of interest are given in Figure 3. Besides the posterior of σ, we present the correlation at distance h, *i.e.* $\mathrm{Corr}(h) = \mathrm{Corr}(F_x, F_{x+h})$, and the posterior of $\phi = 1/(M+1)$. The latter indicates that the normal centring distribution (with mean zero) is a very inadequate description of the data, as could be expected.

Here we present results obtained with the choice of $t^\star = 0.2$ in the prior for λ. The findings are not very sensitive to the value of $t^\star$. Taking $t^\star = 0.05$ gives virtually the same results, with the only slight differences occurring for $\mathrm{Corr}(h)$.

6.2 Volatility Modelling in Time Series

Now we apply our framework to the modelling of financial time series with changing volatility. The modelling of high-frequency financial data, such as exchange rates and stock prices is heavily influenced by two important stylised facts: empirical tails are often heavier than normal and observed series display volatility clustering, in that large values often appear clustered together in time, suggesting that the volatility changes over time.

Many parametric models have been proposed in order to capture these unusual features including

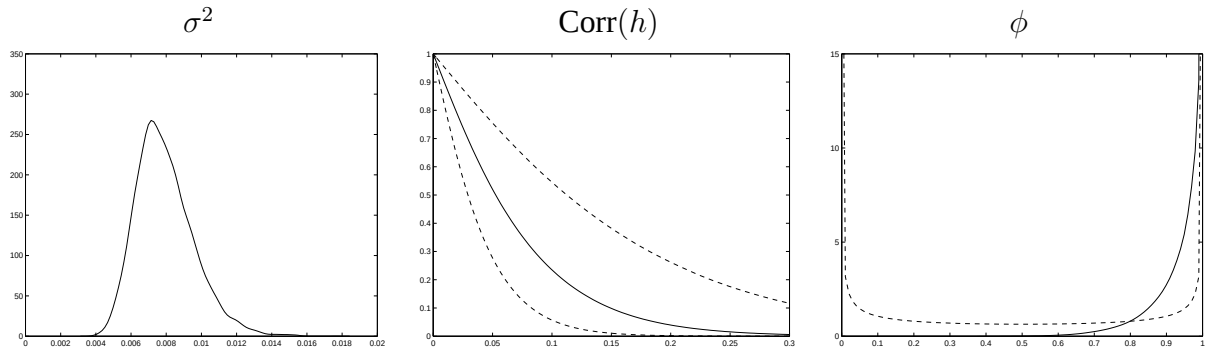

Figure 3: Sine curve regression data: The posterior distributions for the parameter σ^2 and some quantities of interest. The middle panel displays the median (solid line) and the 95% credible intervals (dashed lines) of Corr(h) as a function of h. In the third panel posterior and prior density functions for ϕ are indicated by solid and dashed lines, respectively. We have chosen $t^\star = 0.2$ in the prior for λ.

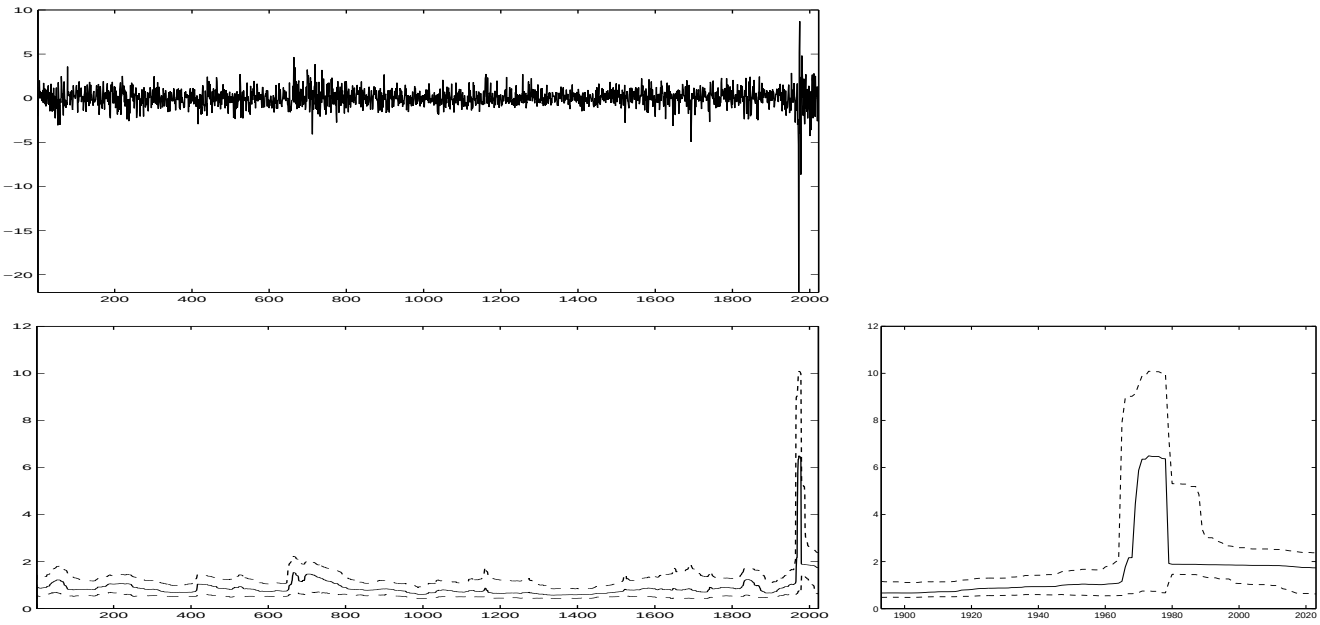

Figure 4: Stock market data: The data on returns are displayed in the top panel. The bottom panels indicate the posterior median (solid lines) and 95% credible intervals (dashed lines) for the volatility distribution F_t. The lower right panel relates to a subset of the data around the 1987 crash. The prior uses the value $t^\star = 100$.

(G)ARCH and stochastic volatility models (see *e.g.* Shephard 1996). A Bayesian semiparametric model is proposed by Kacperczyk *et al.* (2003) who parametrically model the volatility whilst using a Dirichlet process mixture of uniform distributions to model the standardized returns. Jensen (2004) uses a Dirichlet process prior on the wavelet representation of the observables to conduct Bayesian inference in a stochastic volatility model with long memory.

We take the alternative approach to model the volatility through a πDDP, thus inducing time dependence and volatility clustering. In particular, we propose the following discrete-time model where time

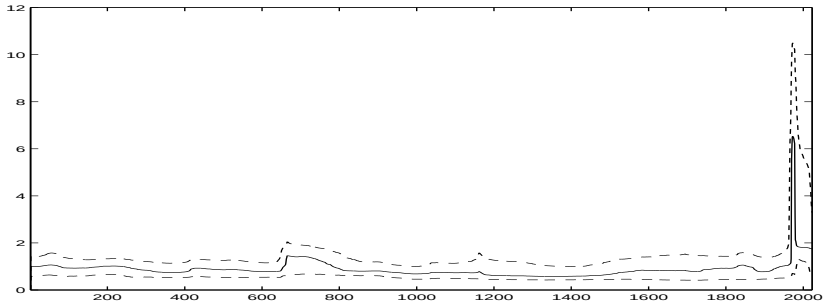

Figure 5: Stock market data: Posterior median (solid lines) and 95% credible intervals (dashed lines) for the volatility distribution F_t. The prior uses the value $t^\star = 300$.

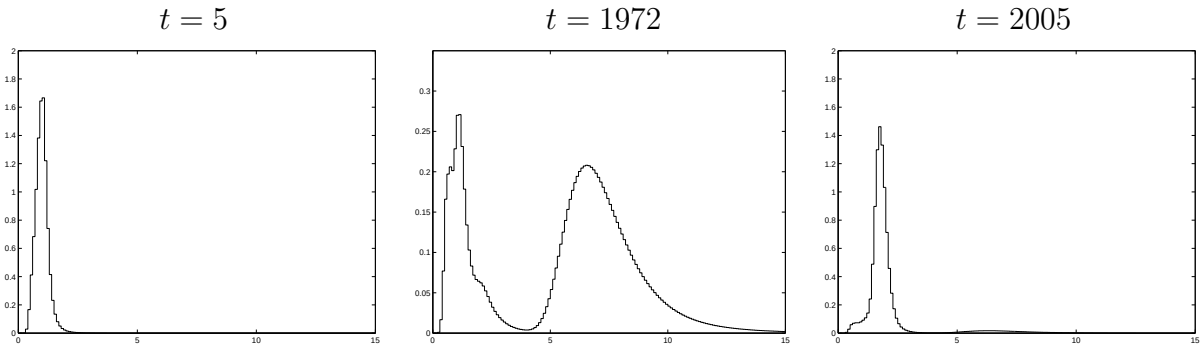

Figure 6: Stock market data: The posterior predictive volatility distribution at various times, using $t^\star = 100$.

$t = 1, \ldots, T$ need not be equally spaced (allowing for possible weekend effects or missing observations):

$$y_t \sim \mathrm{N}(0, \sigma_t^2)$$

$$\sigma_t^2 \sim F_t$$

$$F_t \sim \pi\mathrm{DDP}(MH, \lambda),$$

choosing H to be $\mathrm{IG}(\alpha, \beta)$. We complete the specification with the gamma prior distributions $p(\alpha) = \mathrm{Ga}(0.001, 0.001)$ and $p(\beta) = \mathrm{Ga}(0.001, 0.001)$ and use $t^\star = 100$ and $n_0 = 10$, $\eta = 1$ in the priors for λ and M. The mixture of normals structure of the model will naturally impose heavier than normal tail behaviour. As we are dealing with time series here, we use the arrivals construction to induce the ordering to vary over time.

We use $T = 2023$ daily returns (January 2, 1980 until December 30, 1987) from the Standard and Poor 500 stock price index, displayed in Figure 4 (top panel). The October 19, 1987 crash is immediately obvious from the plot, which also suggests volatility clustering. Sample kurtosis of the returns is 90.3, clearly indicating heavy tails. Figure 4 also tracks the posterior median and the 95% credible interval of the volatility distribution (the time period around the 1987 crash is highlighted in the lower right panel). The flexibility of this modelling of the volatility distribution is apparent: a wide variety of distributions is displayed in Figure 4 and the changes in F_t are quite rapid: the volatility distribution has the potential to change dramatically in a matter of mere days if extreme data events occur. The variety of shapes is

illustrated by Figure 6, where the volatility distributions are plotted at various time points, including the crash date ($t = 1972$). For $t^\star = 300$, as expected, we find that the volatility distributions are somewhat more correlated over time. This leads to a smoother behaviour of the median and credible intervals in Figure 5, which is especially noticeable after the crash date.

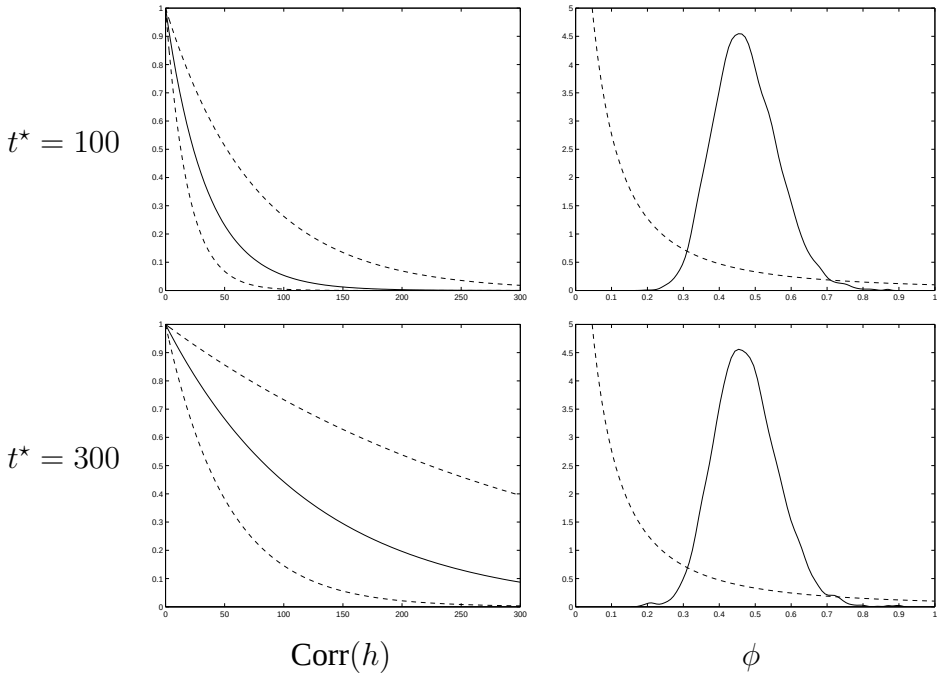

Figure 7: Stock market data: Posterior distributions of the autocorrelation function and $\phi = 1/(M+1)$. In the left panels the solid line are the posterior medians and dashed lines indicate the 95% credible intervals. In the right panels solid lines represent posterior densities and dashed lines priors. The upper panels are for $t^\star = 100$ in the prior for λ and the lower ones correspond to $t^\star = 300$.

More results are presented in Figure 7, where we see confirmation that the autocorrelation of the volatility distribution is somewhat affected by the choice of prior hyperparameter $t^\star$. The inference on ϕ indicates that the inverse gamma centring distribution provides a poor fit to the data.

6.3 Spatial Modelling

An increasingly popular modelling framework for point-referenced spatial data, which originated in geostatistics, is given by

$$y_i = \alpha + \mathbf{f}(\mathbf{x}_i)^T \boldsymbol{\beta} + u_i + \sigma\rho_i, \quad i = 1, \ldots, n, \tag{6}$$

where the mean function $\mathbf{f}(\mathbf{x}_i)$ indicates a known $(p \times 1)$-dimensional vector function of the continuously varying spatial coordinates, with unknown coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$, $(u_1, \ldots, u_n)$ is a realization from a Gaussian process with some spatial correlation structure, and the ρ_i are i.i.d. standard normal, capturing the so-called "nugget effect". The parameter σ is a positive scalar. The Gaussian

assumption on u_i is often considered overly restrictive for practical modelling and a number of more flexible proposals exist in the literature. Of particular relevance for this paper is the nonparametric approach of Gelfand *et al.* (2004), where the locations $\boldsymbol{\theta}$ of the stick-breaking representation of a Dirichlet process are assumed to come from a Gaussian process.

Here we will, instead, use our order-based DDP framework and combine (6) with

$$\alpha + u_i \sim F_{\mathbf{x}_i}$$

$$F_{\mathbf{x}} \sim \pi\text{DDP}(MH, \lambda),$$

where H is a $\text{N}(\mu, \sigma^2/\kappa)$, with $\kappa \sim \text{IG}(0.001, 0.00001)$. The prior distributions assumed for $\boldsymbol{\beta}$ and σ^2 are $\text{N}(0, 1000\sigma^2 I_p)$ and $\text{IG}(0.01, 0.01)$, respectively. The parameter μ is the prior predictive mean of y_i and is chosen to be the sample mean 32.8.

Rather than inducing the dependence through the centring distribution, as in Gelfand *et al.* (2004), we introduce it through similarities in the ordering. Note that we do not need replication, in contrast to the approach of Gelfand *et al.* (2004), and we will use our model on a purely spatial set of temperature data, where only one multivariate observation is available.

In particular, we use the maximum temperatures recorded in an unusually hot week in May 2001 in 63 locations within the Spanish Basque country. As this region is quite mountainous, altitude is added as an extra explanatory variable in the mean function. Throughout, we report results with $t^\star = 2$, which are very close to those obtained with $t^\star = 4$. For the prior on M, we use $n_0 = 1$ and $\eta = 1$.

The main purpose of geostatistical models is prediction, and in Figure 8 we display the posterior predictive distributions at a number of unsampled locations. The lower right panel indicates the location of these unobserved locations (with numbers), as well as the observed ones (with dots). Clearly, there is a variety of predictive shapes with some predictives being multimodal.

Inference on the correlation between distributions at locations that are a distance $t^\star$ apart is given in Figure 9. In comparison with the prior on $\text{Corr}(t^\star)$, which is uniform, the posterior puts less mass on the extremes. The right panel in Figure 9 displays the posterior on ϕ, which indicates that the Gaussian centring distribution is inadequate, but perhaps not dramatically so. Of course, the πDDP mixture model not only allows for departures of the Gaussian model, but also serves to introduce the spatial correlation.

7 Conclusion

We have introduced a framework for nonparametric modelling with dependence on continuous covariates. Starting from the stick-breaking representation we induce dependence in the weights through similarities in the ordering of the atoms. By viewing the atoms as marks in a point process, we implement such orderings through distance measures. Using a Dirichlet stick-breaking representation, we define the class of order-based dependent Dirichlet processes, abbreviated as πDDP's. Observations will update the process locally, in the sense that their effect will vanish as we move further away in the space of the covariates.

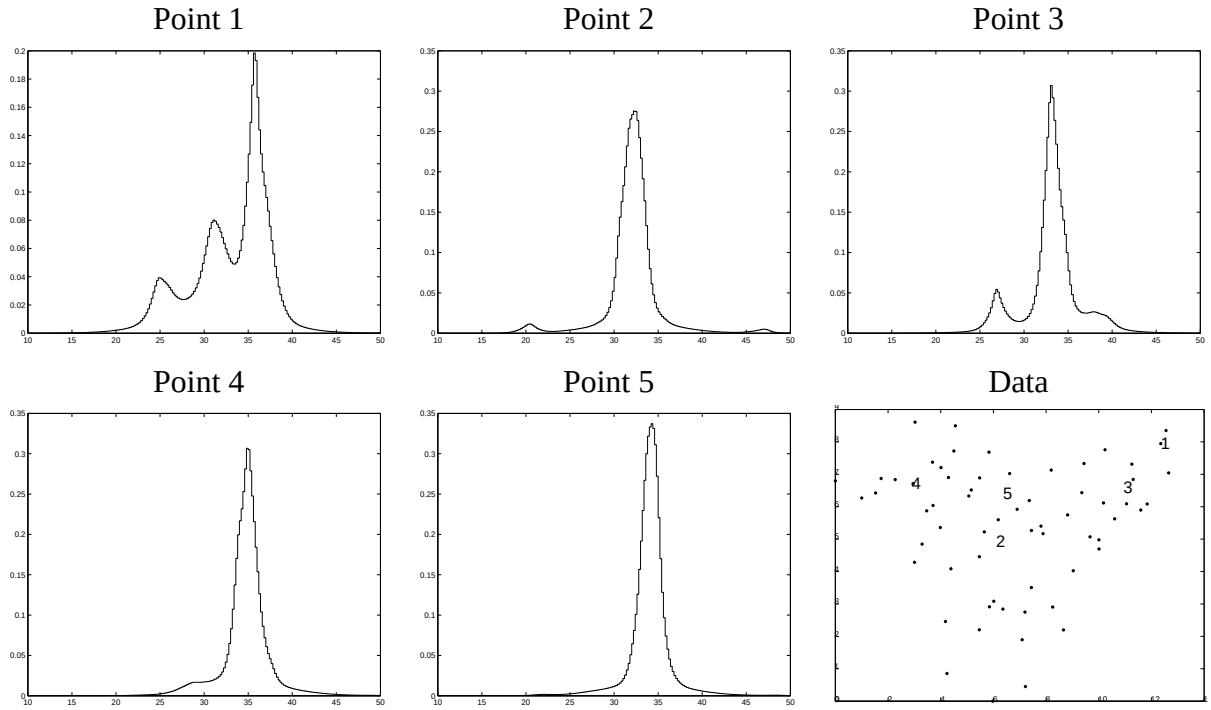

Figure 8: Temperature data: The posterior predictive distribution at five unobserved locations. The latter are indicated by numbers in the lower right-hand panel, where the observed locations are denoted by dots. The prior uses $t^\star = 2$.

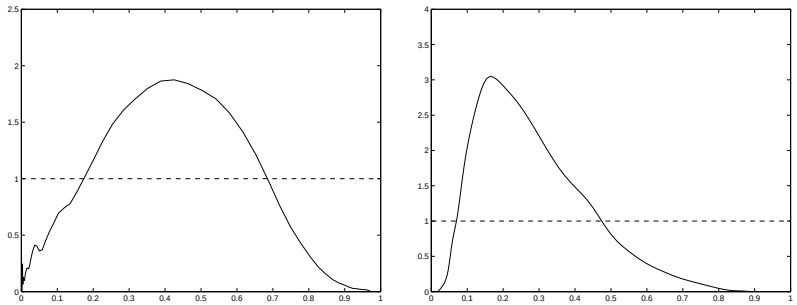

Figure 9: Temperature data: Posterior distributions (solid lines) of the correlation between distributions at distance $t^\star$ and of $\phi = 1/(M+1)$ using $t^\star = 2$. Prior densities are indicated by dashed lines in both panels.

These πDDP's, in combination with Poisson point processes, lead to simple expressions for the correlation function of the distribution, and we propose two specific constructions for inducing an ordering. For mixtures of πDDP's, we design an efficient MCMC sampling algorithm which is able to deal with practically relevant applications.

We apply our framework to a variety of examples: a regression example with simulated data, a stochastic volatility model using a time series of a stock price index, and a spatial model with temperature data. In all cases, the approach using a mixture of πDDP's produces sensible results, without excessive computational effort. We believe the current implementation allows for ample flexibility with-

out requiring very large amounts of data for practically useful inference.

In a wider setting, the basic idea of Order-Based Dependent Stick-Breaking Priors can be used with different marginal stick-breaking priors and different ways of inducing random orderings. The present paper focuses on what we consider a practical implementation, but many other models can be constructed using this or similar frameworks, where *e.g.* we also allow the locations $\boldsymbol{\theta}$ to depend on the covariates.

A Proofs

Proof of Theorem 1

$$\begin{aligned}\mathrm{E}(F_{\mathbf{x}_1}(B)\,F_{\mathbf{x}_2}(B)) &= \mathrm{E}\left[\sum_{i=1}^{n(\mathbf{x}_1)} p_i(\mathbf{x}_1)\delta_{\boldsymbol{\theta}_{\pi_i(\mathbf{x}_1)}}(B)\sum_{j=1}^{n(\mathbf{x}_2)} p_j(\mathbf{x}_2)\delta_{\boldsymbol{\theta}_{\pi_j(\mathbf{x}_2)}}(B)\right]\\ &= \sum_{i=1}^{n(\mathbf{x}_1)}\sum_{j=1}^{n(\mathbf{x}_2)} \mathrm{E}\left[p_i(\mathbf{x}_1)p_j(\mathbf{x}_2)\right]\mathrm{E}\left[\delta_{\boldsymbol{\theta}_{\pi_i(\mathbf{x}_1)}}(B)\delta_{\boldsymbol{\theta}_{\pi_j(\mathbf{x}_2)}}(B)\right].\end{aligned}$$

Now

$$\delta_{\boldsymbol{\theta}_i}(B)\delta_{\boldsymbol{\theta}_j}(B) = \begin{cases} 1 & \boldsymbol{\theta}_i \in B, \boldsymbol{\theta}_j \in B \\ 0 & \text{otherwise}\end{cases}$$

$$\mathrm{E}_{\boldsymbol{\theta}}[\delta_{\boldsymbol{\theta}_i}(B)\delta_{\boldsymbol{\theta}_j}(B)] = \begin{cases} H(B) & i=j \\ (H(B))^2 & \text{otherwise}\end{cases}$$

so that

$$\begin{aligned}\mathrm{E}(F_{\mathbf{x}_1}(B)\,F_{\mathbf{x}_2}(B)) &= (H(B))^2\,\mathrm{E}\left[\sum_{i=1}^{n(\mathbf{x}_1)}\sum_{j=1}^{n(\mathbf{x}_2)} p_i(\mathbf{x}_1)p_j(\mathbf{x}_2)\right]\\ &\quad + \left\{H(B)-(H(B))^2\right\}\sum_{\{(i,j)|\pi_i(\mathbf{x}_1)=\pi_j(\mathbf{x}_2)\}}\mathrm{E}\left[p_i(\mathbf{x}_1)p_j(\mathbf{x}_2)\right]\\ &= (H(B))^2 + \left\{H(B)-(H(B))^2\right\}\sum_{\{(i,j)|\pi_i(\mathbf{x}_1)=\pi_j(\mathbf{x}_2)\}}\mathrm{E}\left[p_i(\mathbf{x}_1)p_j(\mathbf{x}_2)\right],\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}(F_{\mathbf{x}_1}(B), F_{\mathbf{x}_2}(B)) &= \mathrm{E}[F_{\mathbf{x}_1}(B)F_{\mathbf{x}_2}(B)] - \mathrm{E}[F_{\mathbf{x}_1}(B)]\mathrm{E}[F_{\mathbf{x}_2}(B)]\\ &= H(B)(1-H(B))\sum_{k\in T(\mathbf{x}_1,\mathbf{x}_2)}\mathrm{E}\left[V_k^2\right]\prod_{j\in S_k}\mathrm{E}\left[(1-V_j)^2\right]\prod_{j\in S_k'}\mathrm{E}\left[1-V_j\right]\\ &= \frac{2H(B)(1-H(B))}{(M+1)(M+2)}\sum_{k\in T(\mathbf{x}_1,\mathbf{x}_2)}\left(\frac{M}{M+2}\right)^{\#S_k}\left(\frac{M}{M+1}\right)^{\#S_k'}.\end{aligned}$$

Using the form for the variance given in (2), we obtain

$$\mathrm{Corr}(F_{\mathbf{x}_1}(B), F_{\mathbf{x}_2}(B)) = \frac{2}{M+2}\sum_{k\in T(\mathbf{x}_1,\mathbf{x}_2)}\left(\frac{M}{M+2}\right)^{\#S_k}\left(\frac{M}{M+1}\right)^{\#S_k'}.$$

Before proving Theorem 2, we need the following result:

Lemma 1 *For a bounded Borel set B, a stationary Poisson process Φ with intensity λ and $q \in [0,1]$*

$$E_\Phi\left[q^{\Phi(B)}\right] = \exp\left\{-\lambda(1-q)\nu(B)\right\}$$

where ν is the Lesbesgue measure in the appropriate dimension.

Proof: This follows from the definition of the generating functional of a Poisson process. See Stoyan *et al.* (1995, Example 4.2).

Proof of Theorem 2

We need to find the following expectation with respect to the point process:

$$\mathrm{E}_{P_o(\varphi)}\left[\left(\frac{M}{M+2}\right)^{\varphi_{-\mathbf{z}}(S_{-\mathbf{z}}(\mathbf{z}))}\left(\frac{M}{M+1}\right)^{\varphi_{-\mathbf{z}}(S'_{-\mathbf{z}}(\mathbf{z}))}\right].$$

The reduced Palm distribution of a Poisson process is that of a Poisson process with the same intensity (Slivnyak's theorem) and so by Lemma 1 the expectation becomes

$$\exp\left\{-\lambda\frac{2}{M+2}\nu(S_{-\mathbf{z}}(\mathbf{z}))\right\}\exp\left\{-\lambda\frac{1}{M+1}\nu(S'_{-\mathbf{z}}(\mathbf{z}))\right\} = \exp\{-\lambda g(\mathbf{x}_1,\mathbf{x}_2,\mathbf{z})\},$$

where

$$\begin{aligned} g(\mathbf{x}_1,\mathbf{x}_2,\mathbf{z}) &= \left(\frac{2}{M+2}\right)\nu(S_{-\mathbf{z}}(\mathbf{z})) + \left(\frac{1}{M+1}\right)\left[\nu(\{A_1(\mathbf{z})\}_{-\mathbf{z}}) + \nu(\{A_2(\mathbf{z})\}_{-\mathbf{z}}) - 2\nu(S_{-\mathbf{z}}(\mathbf{z}))\right] \\ &= \frac{1}{M+1}\left[\nu(\{A_1(\mathbf{z})\}_{-\mathbf{z}}) + \nu(\{A_2(\mathbf{z})\}_{-\mathbf{z}}) - \frac{2}{M+2}\nu(S_{-\mathbf{z}}(\mathbf{z}))\right], \end{aligned}$$

which directly leads to the result.

Proof of Corollary 1

We consider three different situations. If $z < x_1, x_2$ then

$$\begin{aligned} \{A_1(z)\}_{-z} &= (0, 2(x_1-z)), \quad \{A_2(z)\}_{-z} = (0, 2(x_2-z)), \quad S_{-z}(z) = (0, 2(x_1-z)) \\ \nu(\{A_1(z)\}_{-z}) &= 2(x_1-z), \quad \nu(\{A_2(z)\}_{-z}) = 2(x_2-z), \quad \nu(S_{-z}(z)) = 2(x_1-z). \end{aligned}$$

If $x_1 < z < x_2$,

$$\begin{aligned} \{A_1(z)\}_{-z} &= (2(x_1-z), 0), \quad \{A_2(z)\}_{-z} = (0, 2(x_2-z)), \quad S_{-z} = \emptyset \\ \nu(\{A_1(z)\}_{-z}) &= 2(z-x_1), \quad \nu(\{A_2(z)\}_{-z}) = 2(x_2-z), \quad \nu(S_{-z}) = 0. \end{aligned}$$

If $z > x_1, x_2$,

$$\begin{aligned} \{A_1(z)\}_{-z} &= (2(x_1-z), 0), \quad \{A_2(z)\}_{-z} = (2(x_2-z), 0), \quad S_{-z}(z) = (2(x_2-z), 0) \\ \nu(\{A_1(z)\}_{-z}) &= 2(z-x_1), \nu(\{A_2(z)\}_{-z}) = 2(z-x_2), \quad \nu(S_{-z}) = 2(z-x_2) \end{aligned}$$

The integral in the expression for the correlation function can now be evaluated for the three regions separately

$$\int_{-\infty}^{x_1} \exp\left\{-\frac{\lambda}{M+1}\left[2(x_1-z)+2(x_2-z)-\frac{4}{M+2}(x_1-z)\right]\right\} dz = \frac{M+2}{4\lambda}\exp\left\{-\frac{2\lambda}{(M+1)}(x_2-x_1)\right\}$$

$$\int_{x_1}^{x_2} \exp\left\{\frac{\lambda}{M+1}\left[2(x_1-z)+2(z-x_2)\right]\right\} dz = \exp\left\{-\frac{2\lambda}{M+1}(x_2-x_1)\right\}(x_2-x_1)$$

$$\int_{x_2}^{\infty} \exp\left\{-\frac{\lambda}{M+1}\left[2(z-x_1)+2(z-x_2)-\frac{4}{M+2}(z-x_2)\right]\right\} dz = \frac{M+2}{4\lambda}\exp\left\{-\frac{2\lambda}{(M+1)}(x_2-x_1)\right\},$$

which leads to the result. For $x_1 > x_2$ the proof is analogous.

Proof of Corollary 2

Similar to the proof of Corollary 1. Now, however, the only nonzero integral corresponds to $z < x_1, x_2$, since otherwise $z \notin T(x_1, x_2)$. For this case, we have

$$\{A_1(z)\}_{-z} = (0, x_1 - z), \qquad \{A_2(z)\}_{-z} = (0, x_2 - z)$$

and for $x_1 < x_2$, we get $S_{-z}(z) = (0, x_1 - z)$, which immediately leads to

$$\begin{aligned}\text{Corr}(F_{x_1}, F_{x_2}) &= \frac{2\lambda}{M+2}\int_{-\infty}^{x_1} \exp\left\{-\frac{\lambda}{M+1}\left[(x_1-z)+(x_2-z)-\frac{2}{M+2}(x_1-z)\right]\right\} dz \\ &= \exp\left\{-\frac{\lambda}{M+1}(x_2-x_1)\right\}.\end{aligned}$$

B Correlation functions in higher dimensions with permutations

The 2-dimensional case

For Euclidean distance in 2 dimensions we get

$$\nu(\{A_1(\mathbf{z})\}_{-\mathbf{z}}) = \pi\|\mathbf{x}_1 - \mathbf{z}\|^2$$

$$\nu(\{A_2(\mathbf{z})\}_{-\mathbf{z}}) = \pi\|\mathbf{x}_2 - \mathbf{z}\|^2,$$

and the correlation can be expressed as

$$\text{Corr}(F_{\mathbf{x}_1}, F_{\mathbf{x}_2}) = \frac{4\lambda}{M+2}\int_0^{\pi}\int_0^{\infty} r\exp\left\{-\frac{\pi\lambda}{M+1}\kappa(r,h,\phi)\right\} dr\, d\phi,$$

with

$$\kappa(r,h,\phi) = 2r^2 - 2rh\cos\phi - \frac{2}{\pi(M+2)}\left(r^2\phi + (r^2 - 2rh\cos\phi + h^2)\psi - rh\sin\phi\right) + h^2,$$

where $h = \|\mathbf{x}_1 - \mathbf{x}_2\|$, r and ϕ are the polar coordinates of $\mathbf{z}$, and ψ is defined through $\cos\psi = \frac{h - r\cos\phi}{\sqrt{r^2 - 2rh\cos\phi + h^2}}$.

The d-dimensional case

If we consider $d > 2$ dimensions and again use Euclidean distance $h = \|\mathbf{x}_1 - \mathbf{x}_2\|$, then

$$\text{Corr}(F_{\mathbf{x}_1}, F_{\mathbf{x}_2}) = \frac{2^{d-1}\lambda}{M+2} \int_0^\infty \int_0^\pi \exp\left\{-\frac{\lambda}{M+1} S(r,h,\phi)\right\} r^{d-1} \sin\phi \, d\phi \, dr,$$

with

$$S(r,h,\phi) = \frac{2^{d-1}\pi}{d}(r^d + \hat{r}^d) - \frac{2}{M+2}\left[\frac{2^{d-2}\pi}{d}(r^d(1-\cos\phi) + \hat{r}^d(1-\cos\psi)) - \frac{2^{d-2}\pi}{d(d-1)} h r^{d-1} \sin^{d-1}\phi\right],$$

where $\hat{r} = (r^2 - 2rh\cos\phi + h^2)^{1/2}$ and ψ is as defined above.

References

Antoniak, C. E. (1974): "Mixtures of Dirichlet processes with applications to non-parametric problems," *Journal of the American Statistical Association*, 2, 1152-74.

Bernardo, J.M. and Smith, A.F.M. (1994): "*Bayesian Theory*, Chichester: Wiley.

Carota, C. and Parmigiani, G. (2002): "Semiparametric regression for count data," *Biometrika*, 89, 265-281.

Chib, S. and Hamilton, B. H. (2002), "Semiparametric Bayes Analysis of Longitudinal Data Treatment Models," *Journal of Econometrics*, 110, 67-89.

Chung, Y. S., Dey, D. K. and Jang, J. H. (2002): "Semiparametric hierarchical selection models for Bayesian meta analysis," *Journal of Statistical Computation and Simulation*, 72, 825-839.

Cifarelli, D.M. and Regazzini, E. (1978): "Nonparametric statistical problems under partial exchangeability. The use of associative means," (in Italian) *Annali del'Instituto di Matematica Finianziara dell'Università di Torino, Serie III*, 12, 1-36.

Dahl, D. B. (2003): "An improved merge-split sampler for conjugate Dirichlet Process mixture models," Technical Report 1086, Department of Statistics, University of Wisconsin.

De Iorio, M., Müller, P., Rosner, G.L. and MacEachern, S.N. (2004): "An ANOVA Model for Dependent Random Measures," *Journal of the American Statistical Association*, 99, 205-215.

Denison, D. G. T., Holmes, C. C., Mallick, B. K. and Smith, A. F. M. (2002): "Bayesian Methods for Nonlinear Classification and Regression," Wiley : Chichester.

Doss, D. and Huffer, F. W. (2003): "Monte Carlo methods for Bayesian analysis of survival data using mixtures of Dirichlet process prior," *Journal of Computational and Graphical Statistics*, 12, 282-307.

Escobar, M. D. and West, M. (1995): "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90, 577-588.

Ferguson, T. S. (1973): "A Bayesian analysis of some nonparametric problems," *Annals of Statistics*, 1, 209-30.

Ferreira, J.T.A.S., Denison, D.G.T. and Holmes, C.C. (2002), "Partition Modelling," in *Spatial Cluster Modelling*, eds. A.B. Lawson and D.G.T. Denison, Boca Raton: Chapman-Hall, pp. 125-145.

Gelfand, A.E. and Kottas, A. (2002): "A Computational Approach for Full Nonparametric Bayesian Inference Under Dirichlet Process Mixture Models", *Journal of Computational and Graphical Statistics*, 11, 289-305.

Gelfand, A.E., Kottas, A. and MacEachern, S.N. (2004): "Bayesian Nonparametric Spatial Modelling With Dirichlet Processes Mixing", technical report, Duke University, ISDS.

Ghosh, K., Jammalamadaka, S. R. and Tiwari, R. C. (2003): "Semiparametric Bayesian techniques for problems in circular data," *Journal of Applied Statistics*, 30, 145-161.

Green, P.J. (1995): "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711-732.

Griffin, J. E. and Steel, M. F. J. (2004): "Semiparametric Bayesian Inference for Stochastic Frontier Models," *Journal of Econometrics*, 123, 121-152.

Guidici, P., Mezzetti, M. and Muliere, P. (2003): "Mixtures of products of Dirichlet processes for variable selection in survival analysis" *Journal of Statistical Planning and Inference*, 111, 101-115.

Hansen, M. B. and Lauritzen, S. L. (2002): "Nonparametric Bayes inference for concave distribution functions," *Statistica Neerlandica*, 56, 110-127.

Hirano, K. (2002): "Semiparametric Bayesian inference in autoregressive panel data models," *Econometrica*, 70, 781-799.

Ishwaran, H. and James, L. (2001): "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, 96, 161-73.

Ishwaran, H. and Zarepour, M. (2000): "Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models," *Biometrika*, 87, 371-390.

Jensen, M.J. (2004), "Semiparametric Bayesian Inference of Long-Memory Stochastic Volatility Models," *Journal of Time Series Analysis*, 25. 895-922.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1995): *Continuous Univariate Distributions, Vol. 2*, 2nd. ed., Wiley: New York.

Kacperczyk, M., Damien, P. and Walker, S. G. (2003): "A new class of Bayesian semiparametric models with applications to option pricing," mimeo, University of Michigan Business School.

Knorr-Held, L. and Raßer, G. (2000): "Bayesian Detection of Clusters and Discontinuities in Disease Maps," *Biometrics*, 56, 13-21.

Kottas, A., Branco, M. D. and Gelfand, A. E. (2002): "A nonparametric Bayesian modeling approach for cytogenetic dosimetry," *Biometrics*, 58, 593-600.

Laws, D. J. and O'Hagan, A. (2002): "A hierarchical Bayes model for multilocation auditing," *Journal of the Royal Statistical Society D*, 51, 431-450.

MacEachern, S.N. (1998): "Computational Methods for Mixture of Dirichlet Process Models," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey, P. Muller, D. Sinha, New York: Springer pp. 23-44.

MacEachern, S.N. (1999): "Dependent Nonparametric Processes," in *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA: American Statistical Association.

MacEachern, S.N. (2000): "Dependent Dirichlet Processes," Technical Report, Department of Statistics, Ohio State University.

MacEachern, S.N., Kottas, A. and Gelfand, A.E. (2001): "Spatial Nonparametric Bayesian Models," technical report, Duke University, ISDS.

Mallick, B.K. and Walker, S.G. (1997): "Combining Information From Several Experiments With Nonparametric Priors," *Biometrika*, 84, 697-706.

Medvedovic, M. and Sivaganesan, S. (2002): "Bayesian infinite mixture model based clustering of gene expression profiles," *Bioinformatics*, 18, 1194-1206.

Møller, J. (2003): "Shot noise Cox processes," *Advances in Applied Probability*, 35, 614 - 640

Muliere, P. and Tardella, L. (1998): "Approximating distributions of random functionals of Ferguson-Dirichlet priors," *Canadian Journal of Statistics*, 26, 283-297.

Müller, P., Quintana, F. and Rosner, G. (2004): "A method for combining inference across related nonparametric Bayesian models," in *Journal of the Royal Statistical Society*, Ser. B, 66, 735-749.

Müller, P. and Rosner, G. (1998): "Semiparametric PK/PD models," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, ed. D. Dey, P. Müller and D. Sinha, New York: Springer, pp. 323-337.

O'Hagan, A. and Stevens, J. W. (2003): "Assessing and comparing costs: how robust are the bootstrap and methods based on asymptotic normality?, " *Health Economics*, 12, 33-49.

Papaspiliopoulos, O., Roberts, G, and Sköld, M. (2003): "Non-centred parameterisations for hierarchical models and data augmentation (with discussion)," *Bayesian Statistics 7*.

Papaspiliopoulos, O. and Roberts, G. (2004): "Retrospective MCMC for Dirichlet process hierarchical models," technical report, University of Lancaster.

Pitman, J. (1996): "Some Developments of the Blackwell-MacQueen Urn Scheme," in *Statistics, Probability and Game Theory. Papers in Honor of David Blackwell*, eds. T. S. Ferguson, L. S. Shapeley and J. B. MacQueen, Haywar, California: IMS Lecture Notes - Monograph Series, pp. 245-268.

Sethuraman, J. (1994): "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639-50.

Shephard, N. (1996), "Statistical aspects of ARCH and stochastic volatility," in *Time Series Models in Econometrics, Finance and Other Fields*, eds. D.R. Cox, D.V. Hinkley and O.E. Barndorff-Nielsen, London: Chapman and Hall, pp. 1-67.

Stoyan D., W. S. Kendall and J. Mecke (1995): *Stochastic Geometry and its Applications*, Chichester: Wiley.

van der Merwe, A. J. and Pretorius, A. L. (2003): "Bayesian estimation in animal breeding using the Dirichlet process prior for correlated random effects," *Genetics Selection Evolution*, 35, 137-158.